

**SPATIAL HETEROGENEITY OF RESIDENTIAL LAND
PRICE AND IMPACT FACTORS IN HANGZHOU,
CHINA, 2012 - 2016**

A Thesis

Presented to the Faculty of the Graduate School
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of
Master of Arts in Regional Science

by

Xiao Li

May 2019

© 2019 Xiao Li

ALL RIGHTS RESERVED

ABSTRACT

This research attempts to understand the relationship between different impact factors and the residential land transaction price of Hangzhou, China. Exploiting various online open data sources and conducting detailed data preprocessing, this study builds a spatial econometric model to identify factors which significantly influence the transaction prices over the year 2012 – 2016. Furthermore, the model exams the spatial heterogeneity of different factors across the city. Adopting points of interest (POI) data from Gaode, one of the largest online map service companies in China, the research first defines the centers of different types of locational factors with K-means clustering. The nearest distances to these urban centers are computed accordingly to be set as potential explanatory variables. Collinearity issue is detected and variable selection is modified by variance inflation factors (VIF) calculation and principal component analysis (PCA). From this, land area, plot ratio, distances to major waterbodies, education centers, company centers, government centers, medical centers, tourism centers, opened subway stations and future subway stations which will be opened within three years are selected as independent variables for regression modeling. Finally, the non-stationarity of spatial distributions of estimated coefficients and pseudo t-values of ten explanatory variables indicate that Geographically Weighted Regression (GWR) is more suitable to analyze the spatial variation of impact factors with land price compared to conventional Ordinary Least Square (OLS) regression. The result of GWR model with fixed Gaussian spatial kernels indicates the spatial non-stationary relationship between residential land transaction price and the selected impact factors. The negative impact of most factors is larger in less developed areas than in highly developed areas. With Inversed Distance Weighting (IDW) interpolation improving the data visualization of coefficients spatial distribution, we can identify that land price change is more sensitive to the selective variables in the periphery regions of the city. Distance to job centers shows stronger influence on land price in the south especially the surrounding areas near the information industry

agglomeration identified by K-means clustering. Moreover, two types of subway station factors are both contributing to the land price changes. The stations which will be opened within 3 years are influencing more areas in the city compared to the opened stations, while the magnitude of the coefficient is smaller. It reflects the anticipation of developers about land premium potential. The different spatial patterns of influences from the selected variables on land price implies that urban planners should acquire better understanding about local contexts and conduct location-specific strategies of land price evaluation and land use policies.

BIOGRAPHICAL SKETCH

Xiao Li is a second-year master candidate in Regional Science program at Cornell University. With wide interest in economic geography topics, he is particularly enthused about the application of spatial modeling and urban analytics. He received his bachelor's degree in Resource Management & Urban and Rural Planning in Zhejiang University, China with focus on GIS and regional economic development analysis.

During his two-year study in Cornell, Xiao explored diverse opportunities to further refine his understanding of economic geography theories and applications. In the first year, he combined Location Theory class with Spatial Modeling and Analysis class to provide new information about urban structure of Hangzhou city in China. In the course "Urban Analytics", he developed more skills and tools in information science and benefited a lot from interdisciplinary study. Xiao insists that by embracing the technology innovation with more data and quantification methods, spatial economics study will make greater progress and create new discoveries on classic theories such as Central Place Theory and Bid Rent Theory. He is critically incorporating machine learning into urban economics with specifically careful examination on the concept and construction of smart cities in China.

Xiao is also a person always ready to discover and accept new possibilities and challenges in life. Besides his academic concentration on urban economic studies, he is also interested in urban development history, international finance & trade and urban design. He will keep pursuing his "optimal location" of life in the future with curiosity and perseverance.

**To
Liwei Jia,
The Prettiest Lady,
The Strongest Woman &
The Best Mother in the World**

ACKNOWLEDGMENTS

I would like to first express my gratitude to my advisor Professor Kieran Patrick Donaghy for his encouragement and trust in my entire education at Cornell. His knowledge and wisdom have greatly inspired me to explore diverse opportunities in economic geography study. Although struggling really hard in his Location Theory class, I was able to refine my understanding about classic theories and their evolutions under his patient guidance and instruction. I also want to thank him for holding the regional science group tight and close so every member can exchange and share the ideas and feelings frequently and we all become good friends. Thanks to Professor David Rossiter for his great help and insights in developing this thesis. I had great time discussing progress and obstacles of my work as well as taking his Spatial Modeling and Analysis class. I would also like to thank Professor Nancy Brooks and Professor David Shmoys for their wonderful lectures which aroused my passion on urban topics combining information science and economic theories. I couldn't help but to express how I love the Regional Science program. The uniqueness, compactness and flexibility of Regional Science in Cornell has brought me enormous interdisciplinary exploration combined with my specific research interest.

I want to thank my Regional Science fellows and my dear friends Anna Makido and Edith Zheng Wenyan for always being there for me. We share happiness and sorrow together and with their best support and care, I got through many tough moments. They are not only my classmates; they are my lifelong friends. Thanks to Masaki Kurosaki, Kevin Kimura and Wendy Manyi Lin for spending so many lovely times cooking, swimming and travelling together. Thanks to doctor candidate Shriya Ranjarajan for working together for planning methods class. Thanks to doctor Ziyi Zhang and doctor candidate Yuanshuo Xu for their constructive advice on my thesis and study.

Last but not least, I want to say thanks to my entire family. To my uncles and aunts, my cousins and my grandparents for their unconditional love and support. Thanks to my

parents for their love and understanding. They are the best parents one can ever have. Special thanks to my dear mother, although we always have different opinions, she is forever supportive about my decision.

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH.....	III
DEDICATION	IV
ACKNOWLEDGMENTS.....	V
TABLE OF CONTENTS	VII
LIST OF FIGURES.....	IX
LIST OF TABLES.....	X
LIST OF ABBREVIATIONS	XI
CHAPTER 1.....	1
INTRODUCTION	1
1.1 Background.....	1
1.2 Study Area	1
1.3 Research Objectives.....	4
CHAPTER 2.....	6
LITERATURE REVIEW	6
2.1 Spatial variation of land price.....	7
2.2 GWR studies on land market.....	8
CHAPTER 3.....	12
DATA PREPROCESSING.....	12
3.1 Land transaction data.....	12
3.2 POI data	16
3.3 Subway data.....	17
3.4 Summary of Data.....	19
CHAPTER 4.....	22
METHODOLOGY	22
4.1 K-means clustering and elbow methods.....	22
4.2 VIF analysis	23
4.3 PCA.....	24
4.4 GWR model	25
4.5 IDW interpolation.....	28
CHAPTER 5.....	29
VARIABLES DEFINITIONS AND SELECTION	29
5.1 Definitions of variables.....	29
5.2 VIF & PCA results.....	34
CHAPTER 6.....	43
OLS ANALYSIS	43

CHAPTER 7.....	46
SPATIAL VARIATION ANALYSIS	46
7.1 GWR performance.....	46
7.2 Significance test.....	47
7.3 Interpretation.....	50
7.3.1 Land attributes: land area and plot ratio	53
7.3.2 Pleasant environment: distance to nearest waterbody and distance to nearest tourism center.....	57
7.3.3 Job-housing balance: distance to the nearest company center.....	60
7.3.4 Subway system: stations under operation and stations in the future	62
7.3.5 Summary	67
CHAPTER 8.....	70
DISCUSSION.....	70
8.1 Conclusions.....	70
8.2 Limitations	72
REFERENCES	74

LIST OF FIGURES

FIGURE 1.1: ADMINISTRATIVE DIVISIONS OF HANGZHOU CITY	2
FIGURE 1.2: LAND SELLING OF CHINA'S CITIES IN 2018 (¥ BILLION)	3
FIGURE 1.3: STUDY AREA - MAJOR URBAN AREA OF HANGZHOU.....	4
FIGURE 2.1: BID RENT THEORY MODEL	6
FIGURE 3.1: COORDINATE SYSTEM CONVERSION	14
FIGURE 3.2: METRO SYSTEM IN HANGZHOU CITY	19
FIGURE 5.1: NORMAL LOGARITHM TRANSFORMATION OF DEPENDENT VARIABLE	29
FIGURE 5.2: ELBOW METHOD ANALYSIS FOR DIFFERENT TYPES OF URBAN AMENITIES.....	30
FIGURE 5.3: URBAN CENTERS OF DIFFERENT AMENITIES	31
FIGURE 5.4: BI-PLOT OF PCA WITH 13 VARIABLES	37
FIGURE 5.5: SCATTERPLOT OF VARIABLES DIST_OSUB AND DIST_FSUB.....	42
FIGURE 7.1: DISTRIBUTION OF LOCAL R^2 OF GWR MODEL.....	47
FIGURE 7.2: MAPS OF SIGNIFICANT TEST FOR 10 VARIABLES.....	50
FIGURE 7.3: SPATIAL DISTRIBUTION OF COEFFICIENTS OF IMPACT FACTORS	53
FIGURE 7.4: GWR RESULTS OF THE VARIABLE LAND_AREA.....	55
FIGURE 7.5: GWR RESULTS OF THE VARIABLE PLOT_RATIO.....	56
FIGURE 7.6: GWR RESULTS OF THE VARIABLE DIST_WATER	58
FIGURE 7.7: GWR RESULTS OF THE VARIABLE DIST_TOUR.....	59
FIGURE 7.8: GWR RESULTS OF THE VARIABLE DIST_COP	61
FIGURE 7.9: GWR RESULTS OF THE VARIABLE DIST_OSUB	64
FIGURE 7.10: GWR RESULTS OF THE VARIABLE DIST_FSUB	65

LIST OF TABLES

TABLE 3.1: DATA SUMMARY OF THE LAND TRANSACTION RECORDS OF HANGZHOU	15
TABLE 3.2: 2016 POI DATA OF HANGZHOU FROM GAODE MAP	16
TABLE 5.1 : DATA SUMMARY OF VARIABLES	32
TABLE 5.2: STATISTICAL SUMMARY OF VARIABLES	33
TABLE 5.3: VIF FOR ALL THE VARIABLES	34
TABLE 5.4: IMPORTANCE OF COMPONENTS	35
TABLE 5.5: LOADINGS OF VARIABLES IN EACH COMPONENT	35
TABLE 5.6: VIF RESULTS AFTER VARIABLE SELECTION.....	40
TABLE 5.7: LOADINGS OF SELECTED VARIABLES IN PCA RECALCULATION.....	41
TABLE 6.1: THE RESULT OF COEFFICIENTS OF OLS MODEL.....	44

LIST OF ABBREVIATIONS

GWR	Geographically Weighted Regression
IDW	Inversed Distance Weighting
OLS	Ordinary Least Square
PCA	Principle Component Analysis
POI	Points of Interest
VIF	Variance Inflation Factor

CHAPTER 1

INTRODUCTION

1.1 Background

The Chinese land market has witnessed great booming since 1994, the year when the new tax distribution system was established. With good tax sources being set as national tax and taken by the central government, local governments had to switch their focus to land sales to raise money to alleviate financial burden and support local expenditure. This reliance on land transaction finance system got strengthened after the global financial crisis in 2008, when the Chinese central government decided to drive the economy up by developing the real estate sector. The price of land transaction across the country has hit new highs with an astonishing rate. The land transaction price in China not only reflects the market conditions but also plays as an important economic tool for government regulation. The study on the land price therefore links the land market, urban development and local administration issues together and can provide critical suggestions to policy making and city planning.

1.2 Study Area

The study area is the city of Hangzhou (118°21' - 120°30'E, 29°11' - 30°33'N), a historic but now rapidly developing city in southeastern China. The city is the capital of Zhejiang Province with 10 subordinate county-level districts and 3 counties (see Figure 1.1).

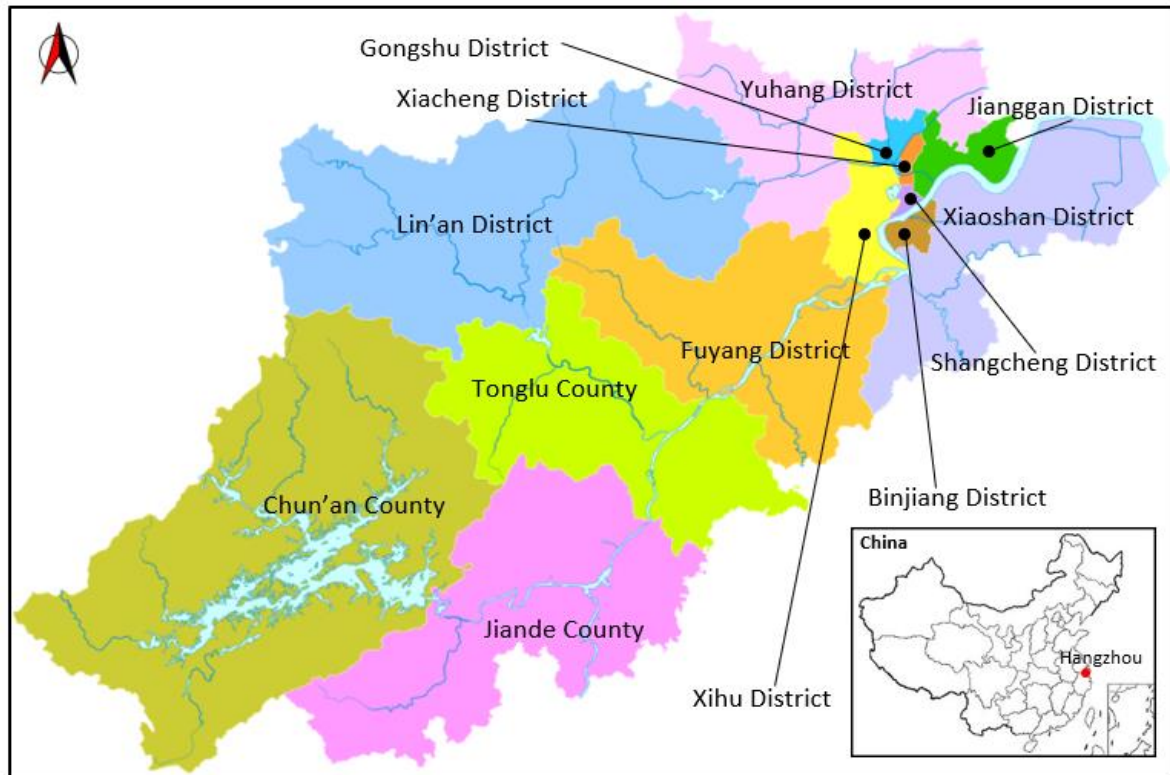


Figure 1.1: Administrative Divisions of Hangzhou City

The main city area consists of Shangcheng district, Xiacheng district, Gongshu district, Binjiang district, Xihu district, Jianggan district, Yuhang District, Xiaoshan District, Lin'an District and Fuyang District, taking up half of the population and one fourth of the total area. Fuyang District and Lin'an District were newly incorporated into the city in 2014 and 2017. In 2018, the total population of Hangzhou is 9.806 million. The main urban area (Shangcheng, Xiacheng, Gongshu, Binjiang, Xihu, Jianggan, Yuhang, Xiaoshan) has more than 70% of the population, reaching 7.235 million¹. The area of Hangzhou and its major urban district are 16853.57 km² and 3352.51 km² in total².

¹ The data is available in the *Population Bulletin of Hangzhou 2018*(in Chinese):

http://www.hangzhou.gov.cn/art/2019/2/14/art_805865_30213801.html

² The area information can be found in the *Hangzhou's First National Geography Census Report*(in Chinese):

<http://www.hzplanning.gov.cn/Data/ResourceFileData/file/20180314/6365662283857022435426970.pdf?WebShieldDRSessionVerify=A91mEZbeEmfIFtoOHeUh>

Steeped in rich history and culture with many human and natural tourism attractions, the city has also adapted itself quickly in the digital era. Hangzhou has the most active private market system and entrepreneurial climate in China. The fast-growing e-commerce industry led by Alibaba has created a new source of economic growth for the city. The city has also won the fame of an emerging technology hub and “Chinese Silicon Valley”. At the same time, Hangzhou is the city whose government heavily relies on land selling revenue for local finance. In 2018, the city ranked the top among all the cities in China for total land transfer fee (Figure 1.2).

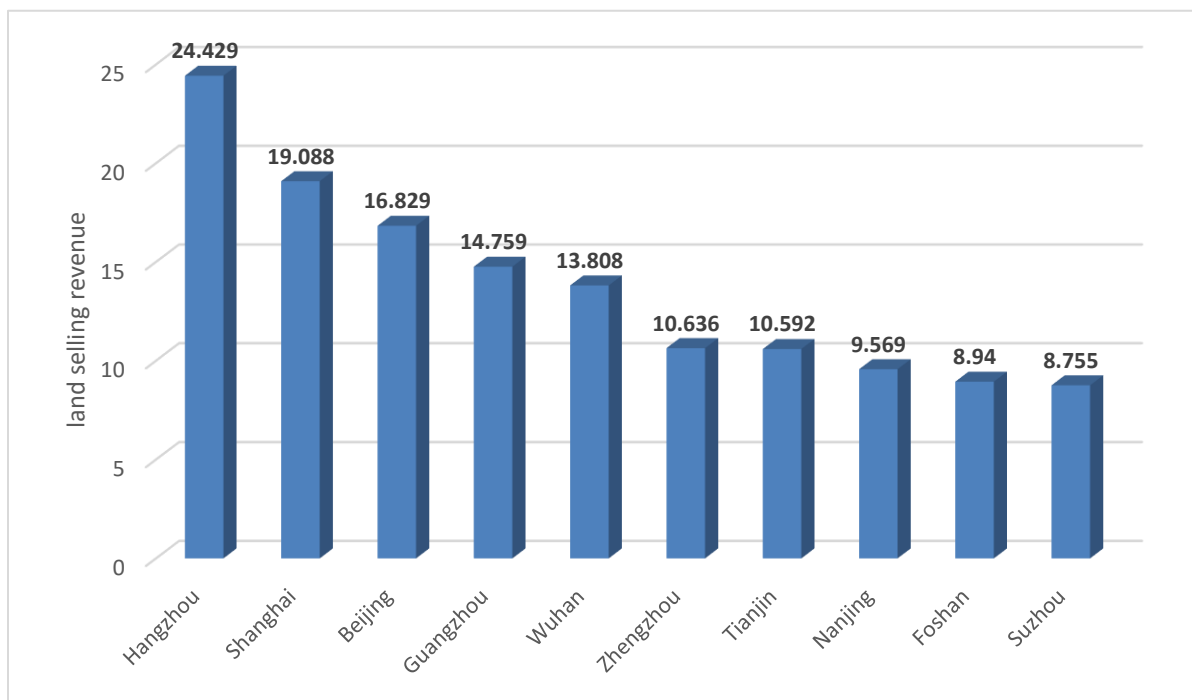


Figure 1.2: Land Selling of China's Cities in 2018 (¥ billion)

Source: CREIS, fdc.fang.com

As the land transaction in the city keeps growing fast, it is important to characterize the spatial distribution of land price, analyze related locational factors and their impact mechanisms. It will help us better understand the changing market, improve the forecast level of urban land price and optimize the land resources allocation.

In this research, I only consider eight main urban districts which have dynamic urban development and detailed land transaction documentation (see Figure 1.3). Hereafter the expression of “Hangzhou” will be referred to the eight main districts (Shangcheng, Xiacheng, Gongshu, Binjiang, Xihu, Jianggan, Yuhang, Xiaoshan)³.

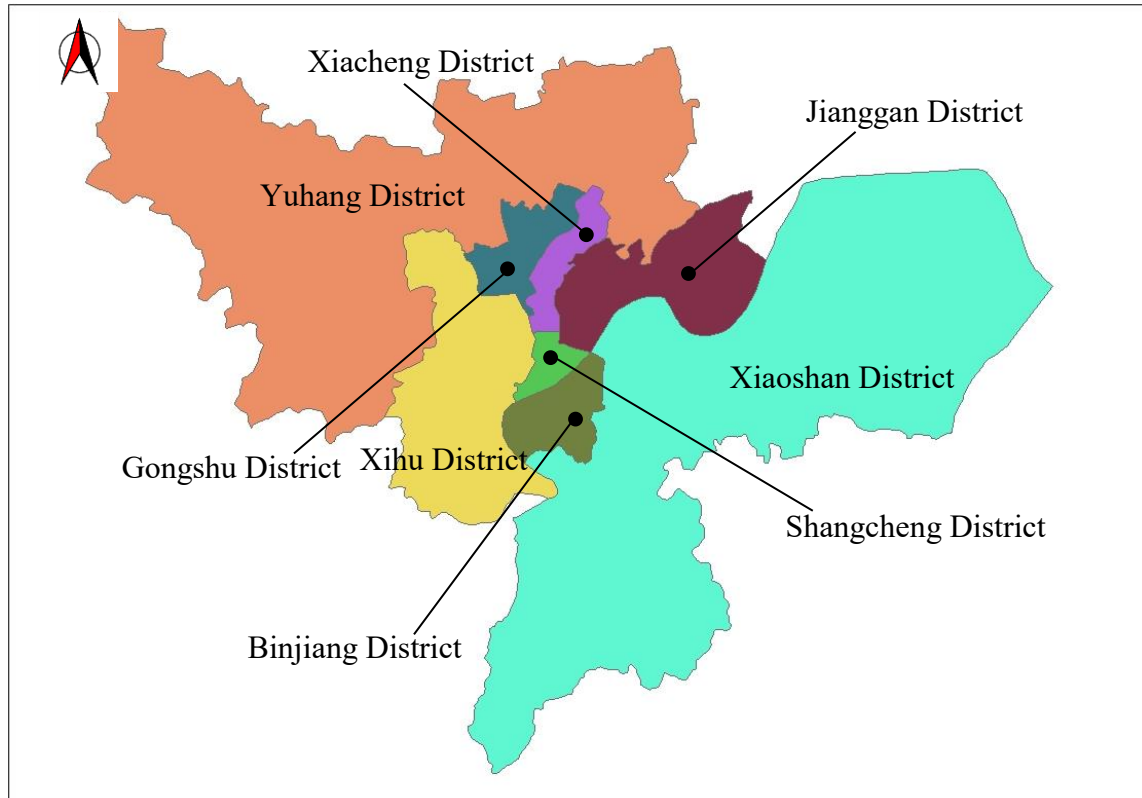


Figure 1.3: Study Area - Major Urban Area of Hangzhou

1.3 Research Objectives

This research aims at identifying the relationship between residential land transaction price and related impact factors in the city of Hangzhou. The following questions and objectives will be the foci for the purpose of the study:

³ For the administrative division data, it is provided in national geographical information monitoring cloud platform website: <http://www.dsac.cn/DataProduct/Detail/201936> (in Chinese)

- 1) How should locational factors be defined in terms of all variety of urban activities, such as education, shopping, working and commuting? Which types of locational factors will influence the land transaction price?
- 2) Given that the subway system will largely reshape the city's landscape, how does it influence the land price? What is the difference of influence between opened stations and planned stations?
- 3) Is there any spatial heterogeneity in these variables? If so, which spatial model should we adopt? What is the interpretation of the modeling results after we take into account the spatial effects?

In order to answer the questions, I first start with discussion of data collection and cleaning in Chapter 3. Locational factors are defined according to the classification of urban activities based on POI data of Hangzhou. The relationship between the subway station and the land transaction price is considered in two forms: (1) the relationship between the opened subway stations and the land transaction price; (2) the relationship between the designed subway stations which will be open within three years and the land transaction price. In Chapter 6 and 7, OLS and GWR models will be applied and results will be compared for spatial analysis. Details of GWR model (kernel, bandwidth, and diagnostics) will also be discussed in Chapter 7. Chapter 8 draws conclusions of the research and extends further discussion on the implications. Limitations of the study are provided for future research improvement.

CHAPTER 2

LITERATURE REVIEW

The study on land price can be traced back to von Thünen, who sought to explain the pattern of agricultural activities surrounding cities in preindustrial Germany (Fujita & Thisse, 2013). Alonso (1964) proposed bid rent theory after the Thünian Model to further explain the bid rent differences among sectors and land price distribution in the city (Figure 2.1). The bid price of the land piece will be higher when its distance to the urban center becomes shorter and transportation cost decreases. Multi-center models were established later and provided richer description and explanation of urban land price structure (Puu, 2012).

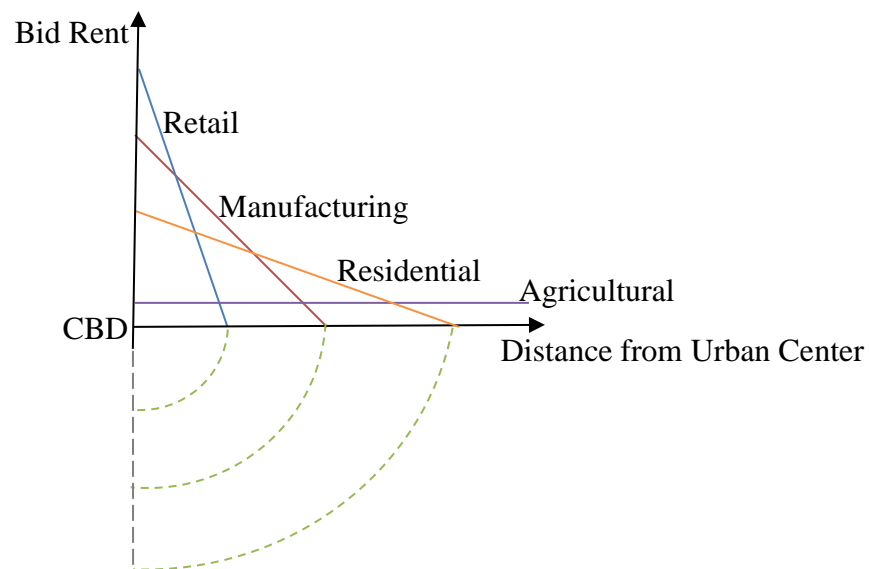


Figure 2.1: Bid Rent Theory Model

Based on these classical theories, the urban land price study has developed rapidly. The Hedonic model (Rosen, 1976) gained popularity for empirical study of city land market as well as housing market. Zheng & Kahn (2008) adopted the hedonic model to analyze the interconnection of the land market and housing market in Beijing, China. The result of the model showed that local public goods, such as access to public transit infrastructure, core high

schools, clean air, and major universities, are important determinants of real estate prices. Qu, et al. (2018) also used the hedonic model with improved feasible generalized least squares method to measure the spatial-temporal differences of the impacts of main driving factors on residential land price of Wuhan, China. Conducting similar research in Nanjing, China, Gao, Chen and Su (2014) took the 96 towns of Nanjing Proper as basic analysis units, and used the aggregate transaction data of residential, industrial, commercial land during 2001-2010 to simulate the main factors influencing the three different types (residential, industrial and commercial land) of land prices with the Hedonic model.

2.1 Spatial variation of land price

Admitting the availability of the traditional Hedonic model, scholars argued that the spatial variation is neglected. While factors are identified using regressions, the spatial attributes of these factors are still unexplored. Spatial non-stationarity needs to be detected and explained for land price study. Spatial econometric analysis is then incorporated to address the issue. Goffette-Nagot et al., (2011) conducted spatial analysis of residential land prices in Belgium. A spatial error model (SEM) and a spatial lag model (LAG) are estimated and compared. The higher likelihood index of the LAG model demonstrated that the data exhibit a spatial pattern in which land prices are influenced by those of neighboring communes instead of by unobserved characteristics in those communes. The results showed that besides classical variables, the linguistic border acted as a strong barrier in the spatial pattern of land prices and that environmental variables have no significant effect in the same linguistic region. In the research of spatial pattern of residential land parcel and its determinants in Beijing (Cui et al., 2017), the spatial error model and spatial lag model were also compared to demonstrate that there indeed exist spatial spillover effects and spatial dependence in residential land price

rather than error dependence. The result of the model showed that the distance to a park or hospital is not influencing the land price. Glumac et al. (2019) adopted a spatial Durbin error model to analyze land transaction prices for Luxembourg between 2010 and 2014 recorded in notarial deeds and cadastral data, together with geo-spatial characteristics. After performing a spatial dependence test and identifying several spatial effects, the spatial model is employed to build up a land price index and can better capture the value of large-scale urban development projects that entail an implicit quality improvement in neighboring areas.

2.2 GWR studies on land market

The study of spatial autocorrelation of land price is associated with the study of spatial heterogeneity. Cao et al., (2013) adopted GWR to explore the effects of various factors on residential land price and their changes in Nanjing major districts. The study compared the differences between the years 2003 and 2009. The initial land price is used as dependent variable instead of transaction price to conduct the modeling. The GWR result demonstrated that the relationships between land price and impact factors vary over space. As the city developed quickly, the variation declined for most of the variables. In the analysis of Wuhan city with 10-year panel data set of residential land price, Hu et al., (2016) found considerable evidence indicating that neighboring land parcels have similar patterns of value visitation and high spatial autocorrelation. By adopting GWR, they found the relationships between land price and impact factors are negative at some locations but positive at other locations. The result showed that in Wuhan city the positive impact of floor area ratio is more significant in highly developed areas than in less developed areas. Conversely, the negative impact of distance to CBD is more significant in highly developed areas than in less developed areas. For the study of city of Hangzhou, Luo (2007) elaborated the algorithm of GWR in his

doctoral dissertation and adopted both fixed kernel and adaptive kernel methods to analyze the land transaction price ranging from 1998 to 2005. The author suggested the result of GWR modeling can be directly applied for revision of land valuation factors and help the government set up the initial price. Zhang (2011) developed mixed GWR in her master thesis, she separated factors into global and local variables to analysis their impact on Hangzhou's land price and achieved a better fitting result.

2.3 Factors affecting land price

Different from Chinese special land ownership regulation, there are fewer literatures on land price from other countries discussing the impact factors using spatial models. But many works focusing on housing price can be helpful especially for factor selection and analysis. Most of researches classify the impact factors into three types: locational factors, neighborhood attributes and land/housing attributes. While there is no clear boundary between locational factors and neighborhood attributes. Some articles also merge locational factors and neighborhood factors together and divide factors into two types. For land attributes, the factors include area, floor area ratio, and green coverage rate (Chau & Chin, 2003). For locational factors, they are often separated into three types: nature amenities: lakes, rivers and mountains; transportation: subway, highway and road density; urban facilities: hospitals, schools, parks and forests (Butler, 1982; Nilsson, 2014; Sui et al., 2015). They also received more attentions compared to the attributes of the land. Jones & Reed (2018) separated urban facilities into active and passive types for the research in Australia. Some studies also took into account the land area and floor area ratio. To quantify locational factors, scholars calculated the distances to the nearest facilities or count the number of facilities within certain buffer of the land parcel. Lv & Zhen (2010) adopted GWR to explore the effects of various

factors on residential land price in Beijing. The quantification method in the research is counting numbers of urban facilities in certain buffers of the land pieces. The study classified the school factor into college level and below-college level to identify the difference and their spatial variations. The model showed that college numbers are more important than below-college level schools for land price. The result also showed that the distance to urban highway has greater influence on land price compared to subway stations. The score of numbers of shopping centers within 1km of the land parcel has no effect on the price. Kan et al., (2019) found that in the year of 2014, employing initial land price data into GWR, the average marginal contribution on the land premium in Nanjing major urban districts from high to low is the distance from CBD, river, expressway, college, hospital, park, primary school and kindergarten. Bus stations and subway stations don't demonstrate any spatial non-stationarity in the model.

While most of literature agreed that the medical, educational and recreational facilities all demonstrate certain degrees of spatial non-stationarity, the relationship of transportation factors and land price or property price always varies from city to city. In Sydney, Australia, the value uplift of housing driven by Bus Rapid Transit (BRT) system is lower than the valuation benefits of BRT in China, Korea and Columbia (Mulley & Tsai, 2016). Dziauddin (2019) adopted a double-log hedonic pricing model combined with GWR and found that the proximity to the nearest light rail transit station gives positive premiums of up to 8% for a majority of properties located in lower-middle and upper-middle income neighborhoods but has a non-significant impact on high-income neighborhoods in Kuala Lumpur, Malaysia. More discussions were involving quantification methods and timing issue of transportation hubs. Nie et al., (2010) identify that metro line 1 construction in Shenzhen, China, has a

negative impact on the price of surrounding properties, while the price will increase sharply 2 years after the operation. Zhang (2011) analyzed Hangzhou's land transaction price with consideration of subway stations opened one year later. In the study of Beijing (Cui et al., 2017), only the opened subway stations are considered to be influential to the price. Kanasugi & Ushijima (2018) examined whether the value of transport innovation is capitalized in land prices immediately after the high-speed railway construction plan is announced in Japan. By clarifying the timing when the benefits of high-speed railway are capitalized in land prices, the result of study implies that benefits are capitalized in land prices when there is demand for time distance shortening immediately after the information disclosure.

2.4 Conclusion

Researches on land price and locational factors have been carried out for a long time and achieved great development on theories and practices. More factors have been considered with various quantification methods and discussed with spatial modeling and analysis. However, many studies neglect the timing issue about factor locations. While normalizing land price data to one specific year over a period, the distance calculations are also based on one-year data of all the facilities. The numbers and locations of different facilities are changing over the years. If the distances are calculated with each individual location of facilities, it is misleading to assume locations and numbers of all the hospitals, schools and parks keep the same over years especially for fast-growing cities. This study will address the issue with data-driven methods and provide better understanding of impact factors built upon previous work.

CHAPTER 3

DATA PREPROCESSING

3.1 Land transaction data

According to government information publicity regulation, a critical government fiscal income source such as land transaction must be documented and published timely. Every local government needs to release all the information of land transactions to the public. Chinese land price monitoring website⁴ co-sponsored by China Academy of Land Survey and Planning and Division of Natural Resources Development and Utilization from Ministry of Natural Resources collects all the information from all the cities and provides detailed information including transaction date, land area, floor area ratio, initial price and land price. Based on the website and local government website information, the China Index Academy⁵ further compiles the data with spatial information of every piece of land. The location information of each land piece is documented with Baidu Map service embedded in the website. The data are provided in Beijing SouFun Science & Technology Development Company website⁶. In this research, I only consider residential land transaction for analysis. The land transaction data of Hangzhou are collected from this website. After data crawling, all the land transaction documents dating back to 2003 are checked to exclude the null information or wrong records.

Both transaction price and initial price were recorded as current price in the dataset, in order to compare the price level among different years for analysis, all the sample prices need to be corrected into normalized prices. To deal with this issue, Hu et al., (2016) adopted the

⁴ <http://www.landvalue.com.cn/> (in Chinese)

⁵ <https://industry.fang.com/en/default.html>

⁶ <https://www1.fang.com/> (in Chinese)

real estate indices to amend the urban land prices to the same date. Lv & Zhen (2010) corrected the land prices of different years by fixed asset investment price index. In this study, due to data accessibility, the price level is readjusted using Consumer Price Index (CPI) from the website of Hangzhou Investigation Team of the National Bureau of Statistics⁷. For regression analysis in the next step, the transaction price is adjusted to unit price per square meter by dividing total transaction price with total floor area. Finally, all the prices are normalized to the level of year 2016 accordingly.

One critical issue with Chinese spatial data is lack of unified coordinate system. Chinese government requires all the online map services to use encrypted geographical coordinate system instead of World Geodetic System (WGS 84) due to national security concern. All the map providers must conduct at least one non-linear coordinate system transformation based on WGS84 before they release the data. Many map service providers such as Gaode and Tectent have adopted GCJ02 (colloquially Mars Coordinates) coordinate system set up by State Bureau of Surveying and Mapping. At the same time, others like Baidu, Inc. developed its own coordinate system BD09 further encrypted upon GCJ02. Spatial information released from different map service providers always has a different coordinate system. In this research, the location information for land transaction is extracted from Baidu online map service embedded in the land transaction website. For distance measurement and calculation, the coordinate system needs to be corrected back to WGS84. The spatial package GeoHey⁸ in QGIS can convert BD09 into WGS84. Applying this package for the data and the location information can be readjusted to WGS84 and the result is displayed in Google Earth for comparisons (Figure 3.1).

⁷ http://www.zjs.gov.cn/hz/zwgk/xxgkml/dcsj_315_1_1/ndsj/201901/t20190109_91476.shtml (in Chinese)

⁸ More information of this package, see: https://plugins.qgis.org/plugins/geohey_toolbox/ (in Chinese)

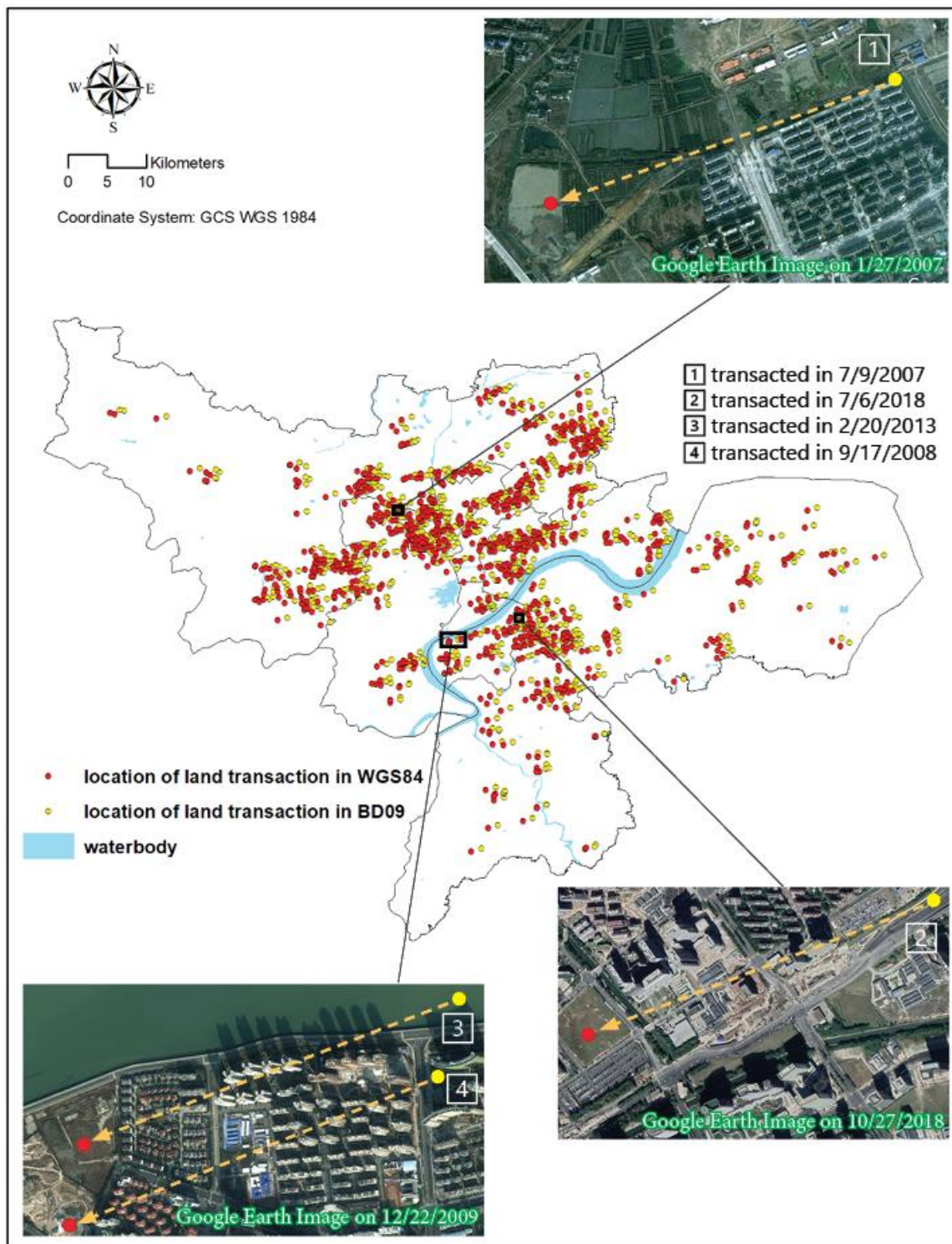


Figure 3.1: Coordinate System Conversion

After cleaning and readjustment, the land transaction data of Hangzhou is well-organized, and all contain information are listed in Table 3.1:

Table 3.1: Data Summary of the Land Transaction Records of Hangzhou

Attribute	Explanation	Unit
Name	The name of the land piece given by the government for land sale record	/
Transaction Date	The date when the land is transacted to the developer	/
Land Area	The total area of the land piece	square meters
Floor Area	The gross area of buildings allowed to be constructed on the land piece	square meters
Plot Ratio	The ratio of total gross floor area divided by the total land area	/
Initial Price	Staring Price settled by government for auction ⁹	ten thousand RMB (¥)
Bid Price	The final transaction price bid by developers	ten thousand RMB (¥)
Premium Rate	Percentage change of the price of the land	/
District	The specific District of Hangzhou where the land piece locates in	/
Longitude_BD	The longitude number extracted from Baidu Map online service	degree
Latitude_BD	The latitude number extracted from Baidu Map online service	degree
Normalized Unit Bid Price	Unit bid price per square meters in the price level of 2016	RMB (¥)/m ²
Longitude_WGS84	The longitude number corrected from BD09 to WGS84	degree
Latitude_WGS84	The latitude number corrected from BD09 to WGS84	degree

⁹ The pricing is settled according to *Rules for Urban Land Valuation (GB/T 18508-2014)*, the regulation is published and implemented in 2014. Source: http://www.mnr.gov.cn/gk/tzgg/201503/t20150320_1991-434.html (The website is in Chinese)

3.2 POI data

Points of Interest (POI) data are point feature data with location information of various types of activities. They can be a gas station, a high school or a fast food chain. Collected by spatial data companies, POI data are easy to access and updated very quickly.

Many map companies provide POI data to public with application programming interface (API). Baidu map provides POI data with restricted number and access, while Gaode map¹⁰ is more open to users and releasing data more often. As one of the biggest online map service providers in China, Gaode has a large client base and developed a complete map service system. The data released from Gaode are complete and reliable for urban studies. Therefore, in this research I choose Gaode Map as my POI data source. The POI data are extracted for Hangzhou in the year 2016 (Table 3.2). Similar to land transaction data, the coordinate system is adjusted to WGS84 from GCJ02 using GeoHey package in QGIS.

Table 3.2: 2016 POI data of Hangzhou from Gaode Map

Categories	Subcategories	Number of Points
Catering	bar, bakery, fast food chain...	9525
Shopping	supermarket, shopping mall, convenience store...	31432
Tourism	park, botanic garden...	1352
Entertainment and Fitness	k-house, theater, fitness center, swimming pool...	4554
Education	college, high school, kindergarten ...	5051
Medical	general hospital, clinic, pharmacy...	5885
Company	factory, office building...	1367
Government	administrative unit, party, court...	5624

¹⁰ For the services provided by Gaode, see the official website for more details: <https://lbs.amap.com/> (in Chinese)

More importantly, POI data contain intensive information of urban activities at high resolution (Yuan et al., 2015). They can reflect both density and diversity of urban activities at the same time. Areas with concentrated different types of POI usually indicate urban clusters or central business districts (CBD). A busy manufacturing zone will be reflected in the online map service with densely distributed POI data in the type of office buildings and factories. Therefore, instead of identifying CBD with preconception, clustering algorithms can be applied with POI data to define different centers of urban area such as working centers in terms of agglomeration of office buildings, education centers in terms of agglomeration of primary schools, high schools and colleges. These centers provide corresponding resources to the city thus create great influence on location choices of people as well as developers. During the bid process developers concern more about the relative location of land parcels to these centers. Individual points of interest may not be influential to the transaction price because they are not stable. Urban centers representing the basic spatial structure of city will keep steady over a period of time, therefore they can generate constant and considerable attraction to people and developers to bid higher for the land close by to utilize the facilities with lower transportation cost.

In this study, instead of using prior knowledge of CBDs and other urban facilities, POI data are analyzed using K-means clustering to identify urban centers of eight types of amenities classified in the dataset. The distances to these centers will be calculated accordingly to be set up as independent variables to evaluate the influence of locational factors to the land price.

3.3 Subway data

Different from urban centers generated from POI data, to identify the influence of subway system to land price, distance to each station is calculated for the measurement. The reason is that the subway stations are all in a connected system. Approaching to one station stands for approaching to the entire public transportation system. Also, as stated in the literature review chapter, the temporal issue of the subway influence needs to be clarified. Besides the current opened subway stations which provide convenient transport networks to the city, future subway stations are also important factors for developers to evaluate the bid price of surrounding land parcels. They represent the premium potential of land purchase therefore are crucial to be considered as impact factors of land transaction prices.

In this research, in order to measure the influence of future subway stations to land bid price and consider the real estate development and the subway construction durations, distance from each land parcel to the nearest subway station which will be opened within three years is calculated. The operation schedule of Hangzhou metro system is provided on the its official website¹¹ and additional information can also be found in corresponding Wikipedia pages¹². For each land piece, stations are classified into “opened” and “future within 3 years” groups according to the comparison of operation schedules and the transaction date. The influences of opened subway stations and future subway stations opened within three years can be analyzed and compared separately. For stations which will be opened in the year of 2019, due to no official announcement on operation date, this research will assume the stations will be opened in the middle of the year. For geospatial data of metro system in Hangzhou, a metro

¹¹ http://www.hzmetro.com/about_1.aspx (in Chinese)

¹² https://en.wikipedia.org/wiki/Hangzhou_Metro and <https://zh.wikipedia.org/wiki/%E6%9D%AD%E5%B7%9E%E5%9C%B0%E9%93%81> (in Chinese)

planning document¹³ and Wikipedia page are adopted to collect and digitize the point data of subway stations (Figure 3.2).

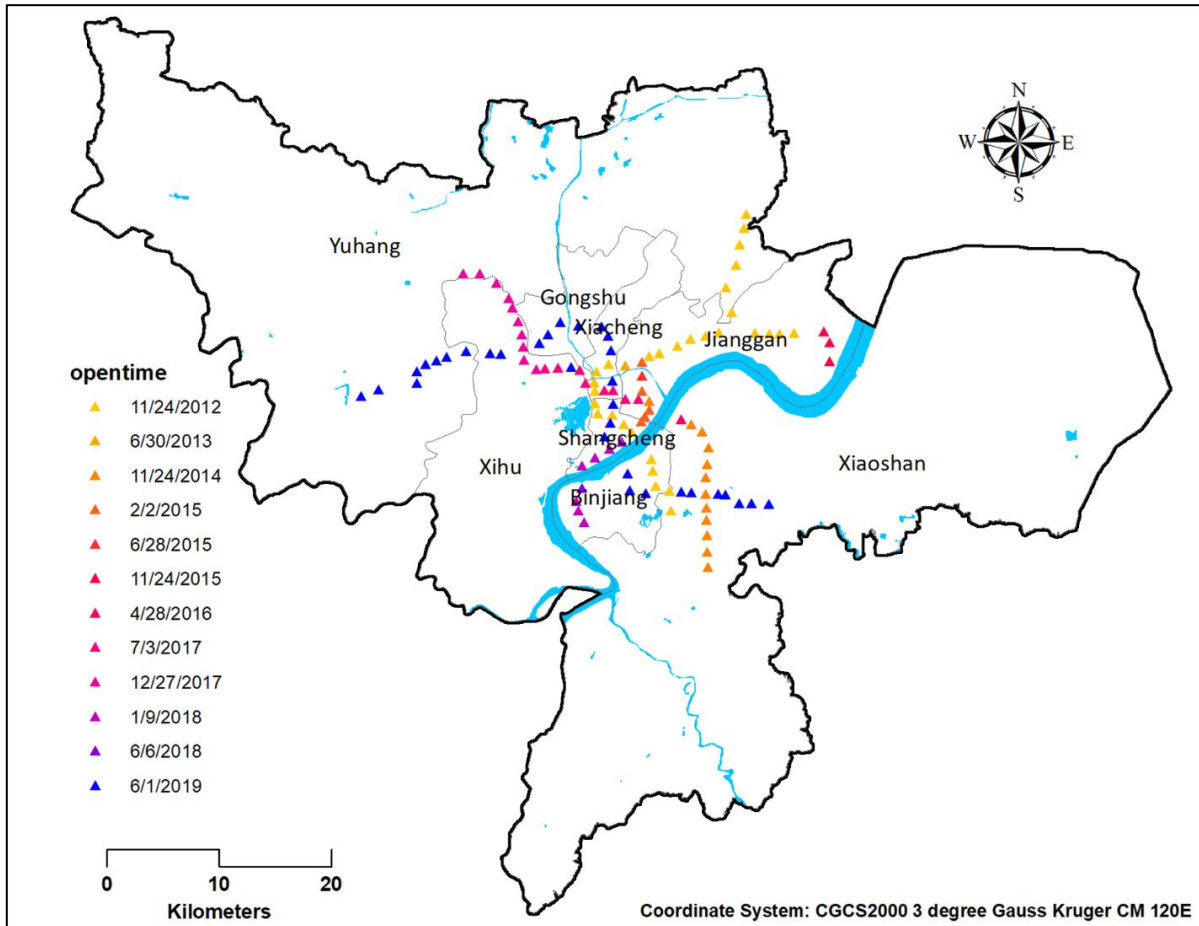


Figure 3.2: Metro System in Hangzhou City

3.4 Summary of Data

While measuring the influence of impact factors by distance between the location of land parcel and the location of individual urban facility (hospital, college, park...), most of Chinese studies don't mention the possibility of emerging of new point (establishment of a new school) or vanishing of old point (relocation of a hospital) when they conduct distance

¹³ The City Metro Construction Plan PhaseIII(2017-2022) can be found in the website of National Development and Reform Commission(in Chinese): <http://www.ndrc.gov.cn/zcfb/zcfbghwb/201612/W020161222571397802300.pdf>

computation. They just assumed that every year the numbers and locations of public facilities won't change. It is acceptable for a short time interval (1 or 2 years) or they can provide evidence that facilities didn't change during the study years, but given the fact that most of cities in China are changing rapidly, it is less reliable to compare year by year data of land transactions with data points of locational factors in a fixed year.

Instead of mapping out individual points of facilities and conducting calculation, this study defines urban centers of different types of amenities and the distances from land parcels to these centers are computed. The advantages are threefold. (1) Compared with single data points, centers representing the basic urban structure are more stable thus can be compared with land transactions in different years (2) For land developers, closeness to stable and resource-rich urban centers can produce more predictable and greater premium of land value and it can further attract more housing consumers and create bigger profit. So, for the bidding process, these centers will generate more influence on the final price (3) Applying data-driven methods to identify urban centers with POI data reduces subjective data points selection based on experience and increase the accessibility of data. The yearly data of individual points of schools, parks or hospitals are difficult to acquire, and the definition and classification of amenities have to be done by large amount of subjective experience.

Based on comprehensive evaluation of data quality and availability, this study will only focus on the period of 2012-2016. Reasons are chiefly as follows: (1) During 2012-2016, the urban planning of Hangzhou didn't change, and the national-level policy stayed stable, so the external influence from policy side can be excluded (2) Hangzhou got its first subway system operating in 2012, in order to calculate the distances from land parcels to opened stations, data points before 2012 cannot be applied. (3) During 2012-2016, the urban planning

of Hangzhou didn't change, and the urban structure remained stable. Although there might be lots of POI emerging or vanishing, the basic structure didn't change, and urban centers also stayed stable. Therefore, distance calculation is based on centers and land parcels are reliable and better characterizing the bid price during the period. After 2016, Hangzhou revised its plan and redesigned its basic structure¹⁴. The urban structure started to change after year 2016. The urban structure generated from 2016 POI data can no longer be representative for the urban landscape thus the data after 2016 cannot be applied. In conclusion, this study will focus on land transaction analysis ranging from 2012 to 2016.

¹⁴ The master plans of Hangzhou city are provided in the website of Hangzhou Urban Planning Bureau: <http://www.hzplanning.gov.cn/hzzg/index.aspx> (in Chinese)

CHAPTER 4

METHODOLOGY

4.1 K-means clustering and elbow methods

K-means clustering is an unsupervised algorithm that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The algorithm clusters data by trying to separate samples in n groups of equal variances, minimizing a criterion known as the inertia or within-cluster sum-of-squares:

$$\sum_{i=0}^n \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

The algorithm divides a set of n samples x into k disjoint clusters C , each described by the mean μ_j of the samples in the cluster. The means are commonly called the cluster “centroids”¹⁵. In this research, different types of POI data are clustered by this algorithm. The centroids generated from K-means clustering are thus defined as centers of corresponding urban activities. What is noteworthy is that centroids are not, in general, points from x , although they live in the same space. Urban centers identified from the clustering algorithm don’t have to be existing points of interest, they reflect the space where the resources are densely concentrated thus creating attraction to developers to offer high land price for land nearby.

The elbow method is a method of interpretation and validation of consistency within cluster analysis designed to help find the appropriate number of clusters in a dataset. This method looks at the total sum of within-cluster sum-of-squares as a function of the

¹⁵The explanation of equations can be found in: <https://scikit-learn.org/stable/modules/clustering.html#k-means>

number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, plotting the sum of within-cluster sum-of-squares against the number of clusters, the first clusters will add much information (explain a lot of variance), if at some point the marginal gain drops, giving an angle in the graph. The number of clusters is chosen at this point, hence the “elbow criterion” (Ketchen & Shook, 1996).

The combination of K-means clustering and elbow method with POI data will identify both numbers and locations of different urban centers. The distances are calculated accordingly to measure their influence on land transaction price.

4.2 VIF analysis

Variance inflation factor (VIF) is the ratio of variance in a model with multiple terms divided by the variance of a model with one term alone (Allison, 1999). It quantifies the severity of multicollinearity in an ordinary least square regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

For m independent variables in a linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_m x_m + \varepsilon$$

The ordinary least square regression is conducted with x_j as a function of all the other explanatory variables:

$$x_j = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{j-1} x_{j-1} + \alpha_{j+1} x_{j+1} + \cdots + \alpha_m x_m + \varepsilon$$

where α_0 is a constant and ε is the error term. Calculation of VIF factor for variable x_j then can be expressed as:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination of the regression equation above, with x_j on the left-hand side, and all other independent variables on the right-hand side.

VIF for each individual predictor can be generated with the method. The magnitude of multicollinearity is analyzed by the size of VIF. Generally speaking, if VIF is greater than 10, the collinearity is high. For many researches, 5 or 8 is also considered to be set up as the threshold of strong collinearity. This method is applied in this study to detect multicollinearity issue among variables.

4.3 PCA

PCA is a statistical procedure for dimension reduction and data visualization (James, 2013). It uses orthogonal transformation to convert a set of observations of correlated variables into a set of values of linearly uncorrelated variables. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set (Hotelling, 1933). Therefore, it is a useful tool for solving the collinear issue.

Furthermore, in regression analysis, the larger the number of explanatory variables, the greater is the chance of overfitting the model, producing conclusions that fail to generalize to other datasets. PCA can also help reduce them to a few principal components especially when there are strong correlations between explanatory variables. The dimension reduction

can be realized by calculating the proportion of variance explained for each component and setting up a threshold to pick components.

One limitation of PCA is that after transformation, there is no guarantee that the orthogonal dimensions are interpretable, new variables may not have any social or economic meaning. In this study, PCA will be applied after VIF analysis, the collinearity discovered by VIF will be addressed by PCA with analysis of loadings (eigenvectors) and biplots. Most representative variables will be selected accordingly.

4.4 GWR model

Geographically Weighted Regression (GWR) is an extension of linear or generalized linear regression, which re-fits the regression equation at each data point, based on some neighborhood and weighting scheme. A main use of GWR is to detect if there is spatial non-stationarity in the linear model. A global model only representing the overall relation may miss critical local variations (Brunsdon, Fotheringham & Charlton, 1996).

GWR is a locally adaptive version of ordinary linear regression applied over the entire point set, it can be expressed as follows:

$$y_i = \beta_{i0} + \sum_{l=1}^p \beta_{il}x_{il} + \varepsilon_i$$

where y_i is the dependent variable at location i , x_{il} is the l th explanatory variable at location i , β_{i0} is the intercept parameter at location i , β_{il} is the coefficient parameter of l th explanatory variable at location i , and ε_i is the random disturbance (Tamesue & Tsutsumi, 2013).

The coefficients are estimated via weighted least squares, where each observation is weighted according to the distance from location i , and matrix notation of the estimators is expressed as:

$$\beta(i) = (\mathbf{X}^T \mathbf{W}(i) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(i) \mathbf{y}$$

The \mathbf{X} matrix includes all points as in the global model, $\mathbf{W}(i)$ is the diagonal matrix containing the weights of the points in the neighborhood to be used to fit the regression for location i :

$$\mathbf{W}(i) = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix}$$

The diagonals are calculated based on a given spatial kernel function. There are various choices for the kernel functions, such as the bi-square function, the Gaussian function, and the exponential function. The bi-square function and the Gaussian function are two commonly used kernel functions, and their expressions are listed as following:

$$\text{Gaussian function: } w_{ij} = e^{-\frac{1}{2}(\frac{d_{ij}}{h})^2}$$

$$\text{Bi-square function: } w_{ij} = (1 - (\frac{d_{ij}}{h^2})^2)^2 \text{ if } d_{ij} \leq h, \text{ else } w_{ij} = 0$$

Two kernel functions have different implications: for bi-square function, it assumes that beyond a certain threshold distance (bandwidth) h , samples create no contribution to the local regression model. However, Gaussian method adopts a smooth function weighting all data samples. Points beyond the bandwidth are still considered in the modeling process, but the corresponding weights decrease sharply according to the function.

In the micro-level study of city scale, land transactions are influenced by all other samples. To better reflect the distance decay principle of bid rent theory over the city, it is

more reasonable to use Gaussian function. The GWR model in this research is built upon Gaussian kernel function.

Similar to kernel function selection, to define the optimal bandwidth h , there are also two methods to choose: (1) fixed method; (2) adaptive method. For fixed bandwidth method, it uses the distance parameter h in the above formulation, and only considers points within that distance (bi-square kernel) or decay according to it (Gaussian kernel). As for adaptive method, it considers a proportion of the points to use for each local fit, and then weights them according to the kernel function. It ensures that there are enough points with sufficient weight to calibrate the regression. The land transaction records in Hangzhou are spatially unevenly distributed, some land parcels are far away from the rest of samples, if the adaptive method is adopted, these parcels are modeled with remote samples to ensure enough proportion of points. However, samples which are too far away don't have influence on these parcels. The modeling process may generate results which are not following distance decay rule and will lead to incorrect interpretation of spatial variation. Therefore, this study adopts fixed bandwidth method combined with Gaussian kernel to set up the GWR model.

To determine the optimal bandwidth, this research adopts Cross Validation (CV) method, by which the bandwidth is estimated by minimizing the CV as follows:

$$CV = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_{\neq i}(h))^2}$$

where $\hat{y}_{\neq i}(h)$ is a predictor at location i using data excluding this local point, n is the total number of points. In other words, it computes the GWR with the proposed bandwidth using all the points, then applies it at known points without considering that point and collects the errors. Optimal bandwidth thus can be identified by minimization of CV.

In summary, this study will apply the fixed bandwidth GWR with Gaussian spatial kernel to the land price dataset. The result will be analyzed by a significance test and the spatial pattern discovered by the model will be explained.

4.5 IDW interpolation

To better visualize the results generated from the GWR model, a simple Inversed Distance Weighting (IDW) method is adopted to interpolate the entire region based on the data points of actual land transaction. This method is built upon the assumption that things close to one another are more alike than those farther apart, each known point has a local impact that declines with distance. To predict a value for any unmeasured location, IDW uses the measured values surrounding the prediction location by assigning them with different weights according to the distances.

The mapping results are not to intended produce predictions of local coefficients over the space but to provide a more general spatial trend visualization. In Chapter 7, the spatial distributions of coefficients are interpolated with this method implemented in ArcGIS¹⁶. IDW in the software relies mainly on the inverse of the distance raised to a mathematical power and controls the number of input points to produce the results. In this research, because we only use this method for visualization, we use the default IDW setting with power of 2 and variable search radius with fixed number of input points. The explanation and interpretation of the spatial distributions are provided by combining the IDW results and the maps of coefficients of actual transaction points.

¹⁶ The data visualization with IDW is realized in ArcGIS 10.5, for the detailed information of how it works: <http://desktop.arcgis.com/en/arcmap/latest/tools/3d-analyst-toolbox/idw.htm>

CHAPTER 5

VARIABLES DEFINITIONS AND SELECTION

5.1 Definitions of variables

To understand how the locational factors influence the land price and their spatial variation, this research adopts distance calculation to measure the impact of factors to transaction price. The hypothesis is that the residential land transaction price is higher when its location is closer to certain types of urban centers. Also, for linear regression analysis, the dependent variable is required to be normally distributed, therefore the land transaction price is transferred into the normal logarithmic value and the result is shown in Figure 5.1:

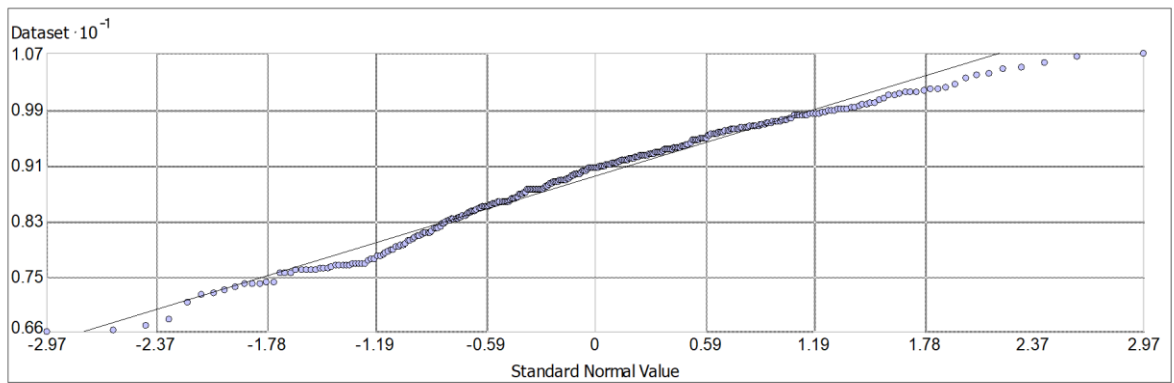


Figure 5.1: Normal Logarithm Transformation of Dependent Variable

For locational factors, K-means clustering illustrated the Chapter 4.1 is applied to the POI dataset. Eight types of urban facilities are analyzed with elbow method to identify optimal k indicating the number of centers in the city. The process is implemented in Python with scikit-learn package¹⁷. Plots are demonstrated as follows:

¹⁷ See the official website for detailed information of the package: <https://scikit-learn.org/stable/>

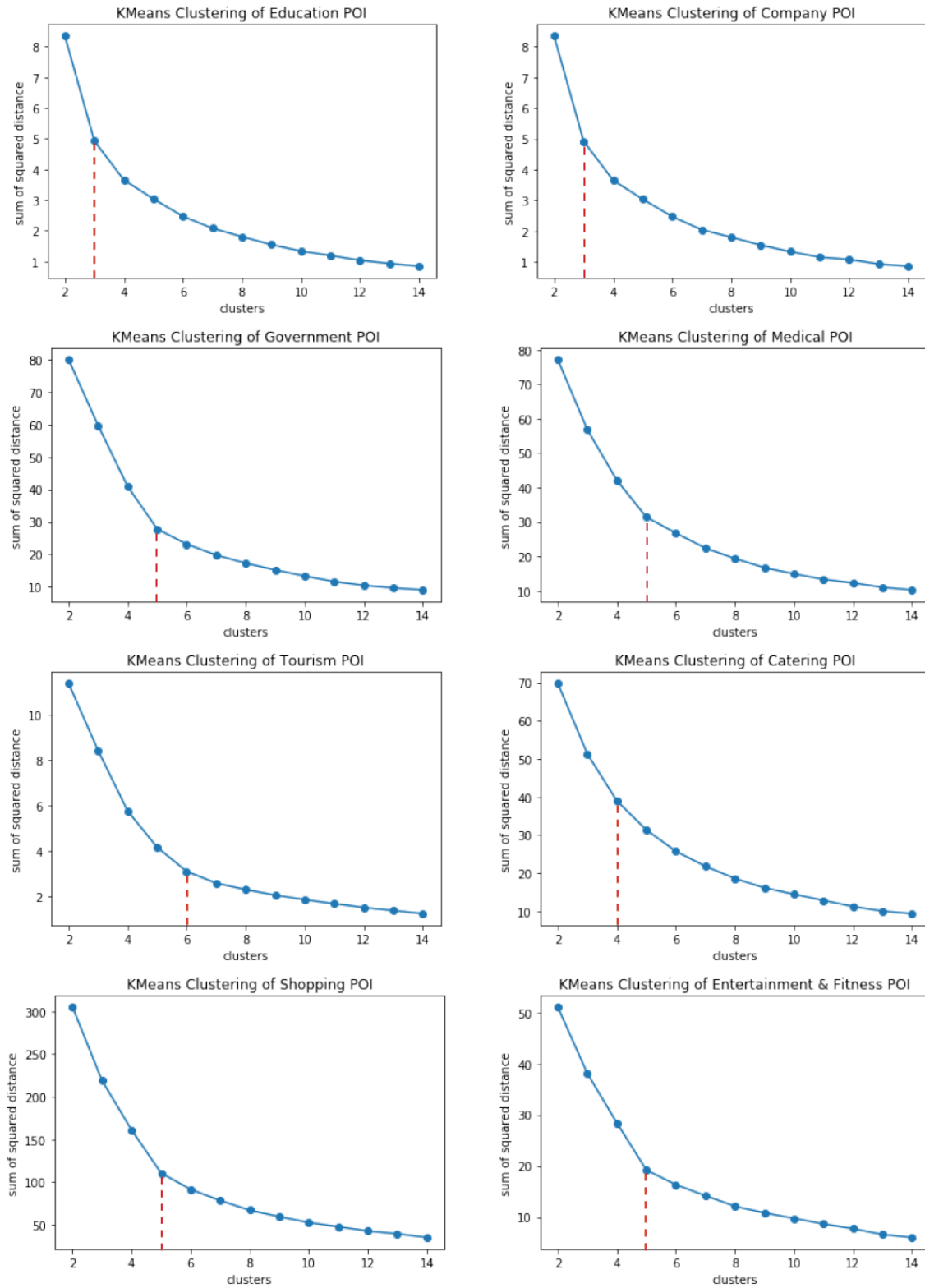


Figure 5.2: Elbow Method Analysis for Different Types of Urban Amenities

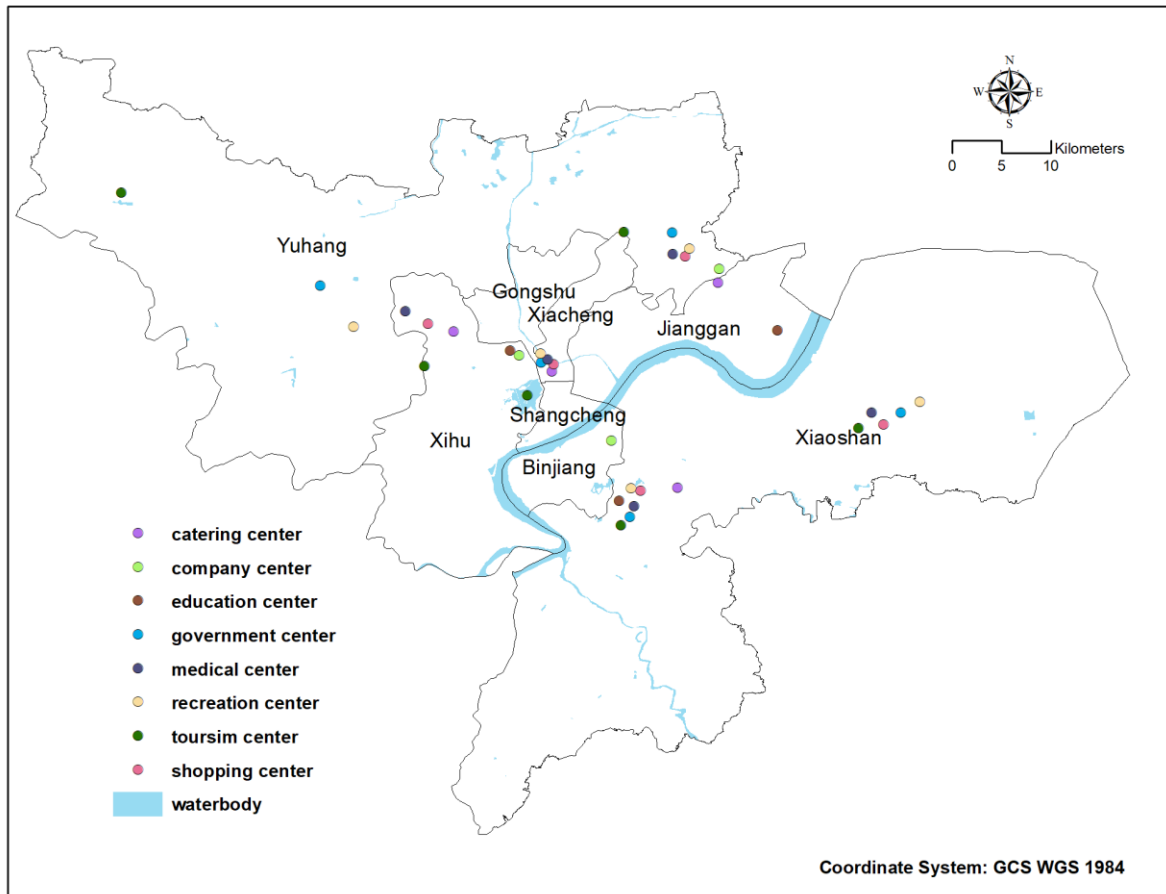


Figure 5.3: Urban Centers of Different Amenities

The distances calculation for subway system and waterbody are implemented in ArcGIS based on the project coordinate system CGCS2000/3-degree Gauss-Kruger CM 120E. All the variables are listed in the table as follows:

Table 5.1 : Data Summary of Variables

Variable Definition	Abbreviation	Geometrical Features
Independent Variables		
Land Attributes		
Land area	land_area	-
Plot ratio	plot_ratio	-
Locational Factors		
Distance to the nearest education center	dist_edu	Point
Distance to the nearest company center	dist_cop	Point
Distance to the nearest government center	dist_gov	Point
Distance to the nearest medical center	dist_med	Point
Distance to the nearest tourism center	dist_tour	Point
Distance to the nearest catering center	dist_cater	Point
Distance to the nearest shopping center	dist_shop	Point
Distance to the nearest entertainment & fitness Center	dist_etm	Point
Distance to the nearest waterbody	dist_water	Polygon
Distance to the nearest opened subway station	dist_osub	Point
Distance to the nearest future subway station opened within 3 years	dist_fsub	Point
Dependent Variable		
Normal logarithmic value of normalized unit bid price of land transaction	log_nubp	-

The statistical summary of all the variables including normalized unit bid price is displayed in Table 5.2:

Table 5.2: Statistical Summary of Variables															
	Normalized Unit Bid Price	ln_nubp	land_area	plot_rat io	dist_edu	dist_cop	dist_gov	dist_med	dist_tour	dist_cater	dist_shop	dist_etm	dist_wate r	dist_osub	dist_fsub
Unit	<i>RMB (¥)/m²</i>	/	<i>m²</i>	/	<i>km</i>	<i>km</i>	<i>km</i>	<i>km</i>	<i>km</i>	<i>km</i>	<i>km</i>	<i>km</i>	<i>km</i>	<i>km</i>	<i>km</i>
Mean	10075.31	8.92	48072.76	2.41	9.58	9.32	8.02	7.14	7.55	7.82	6.96	7.07	3.42	7.59	8.41
Standard Error	419.07	0.05	1961.65	0.04	0.27	0.31	0.18	0.20	0.17	0.27	0.21	0.20	0.14	0.36	0.37
Median	8524.16	9.05	41060.00	2.40	9.20	8.59	8.08	6.39	7.46	7.00	6.13	6.68	2.60	6.02	7.24
Standard Deviation	7635.76	0.82	35742.94	0.72	4.89	5.74	3.31	3.57	3.09	4.90	3.91	3.56	2.55	6.48	6.73
Minimum	769.80	6.65	455.00	0.18	0.90	1.51	1.04	0.68	0.68	0.92	0.72	0.93	0.09	0.12	0.14
Maximum	45367.86	10.72	293354.00	7.30	39.91	40.83	22.25	29.95	20.65	34.69	32.33	27.03	10.46	43.19	46.54
Sum	3345002.03	2961.58	15960157.00	800.46	3180.62	3094.07	2661.86	2369.95	2506.99	2595.09	2310.30	2347.85	1133.86	2518.64	2791.67
Count	332	332	332	332	332	332	332	332	332	332	332	332	332	332	332

5.2 VIF & PCA results

To analyze the relationship among all the independent variables, VIF is applied to the dataset and the result is as follows:

Table 5.3: VIF for All the Variables	
Dependent Variables	VIF
land_area	1.09
plot_ratio	1.18
dist_edu	4.74
dist_cop	10.63
dist_gov	7.48
dist_med	31.18
dist_tour	1.74
dist_cater	7.98
dist_shop	27.52
dist_etm	7.40
dist_water	2.07
dist_osub	6.11
dist_fsub	2.24

The result shows that there is a strong collinearity issue among variables. A multivariate regression model with collinear predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others. To address this issue, Principal Component Analysis (PCA) is adopted and variable selection is based on the interpretation of PCA result. Table 5.4 shows the importance of components:

Table 5.4: Importance of Components

	Standard deviation	Proportion of Variance	Cumulative Proportion
PC.1	2.588	0.515	0.515
PC.2	1.376	0.145	0.66
PC.3	1.036	0.083	0.743
PC.4	0.975	0.073	0.816
PC.5	0.783	0.047	0.863
PC.6	0.728	0.041	0.904
PC.7	0.627	0.03	0.934
PC.8	0.609	0.029	0.963
PC.9	0.417	0.013	0.976
PC .10	0.372	0.011	0.987
PC.11	0.304	0.007	0.994
PC.12	0.226	0.004	0.998
PC.13	0.133	0.002	1.000

We can see from the result that when the number of principal components reaches eight, the cumulative proportion has already passed 95%. The loading factors of variables in each principal component is displayed in Table 5.5 below:

Table 5.5: Loadings of Variables in Each Component

	PC.1	PC.2	PC.3	PC.4	PC.5	PC.6	PC.7	PC.8	PC.9	PC .10	PC.11	PC.12	PC.13
land_area			-0.74	0.63	-0.20								
plot_ratio		0.13	0.61	0.69		-0.29				-0.11			
dist_edu	-0.32	-0.25		0.14	0.18		0.29		0.72	0.38			
dist_cop	-0.34	-0.2			-0.16	0.17		-0.41		-0.40		-0.65	
dist_gov	-0.31	0.21	-0.10			-0.5	-0.37	0.12		0.14	0.56	-0.11	-0.21
dist_med	-0.35	0.24				-0.11	0.17	0.19	-0.21	0.33	-0.14	-0.22	0.71
dist_tour	-0.17	0.44		0.15	0.53	0.48	-0.45						
dist_cater	-0.34					0.25	0.36	-0.18	-0.40		0.53	0.41	
dist_shop	-0.36	0.16					0.26		-0.32	0.15	-0.41	-0.21	-0.64
dist_etm	-0.32	0.23		-0.16		-0.28	0.18		0.25	-0.65	-0.23	0.35	
dist_water		-0.57	-0.15		0.65	-0.33	-0.11		-0.24	-0.11	-0.13		
dist_osub	-0.31	-0.21			-0.35		-0.48	-0.38	-0.11	0.22	-0.32	0.39	
dist_fsub	-0.25	-0.32	0.15	0.16	-0.23	0.28	-0.21	0.75		-0.15			

The loading factor indicates the contribution of each variable to the newly transformed component. In principal component 1, all the non-nature locational factors are dominant. This

grouping has explained 51.5% of total variance. The variable of distance to waterbody has the largest absolute loading in principal component 2 and land area and plot ratio dominate in principal component 3. We can conclude that land attribute factors are not correlated to locational factors, at the same time, social locational factors demonstrate a certain degree of correlation. To further disentangle correlations among variables, bi-plot is analyzed (Figure 5.4).

The two-dimension figure can demonstrate the collinearity among variables. The directions of arrows indicate the collinearity among factors. The closer the directions, the stronger the collinearity. Distance to shopping centers and distance to medical centers have strong collinearity. Distance to catering centers and distance to company centers also have collinearity. Distance to waterbody and distance to tourism centers are isolated in the graph showing the two factors have less collinearity with other factors.

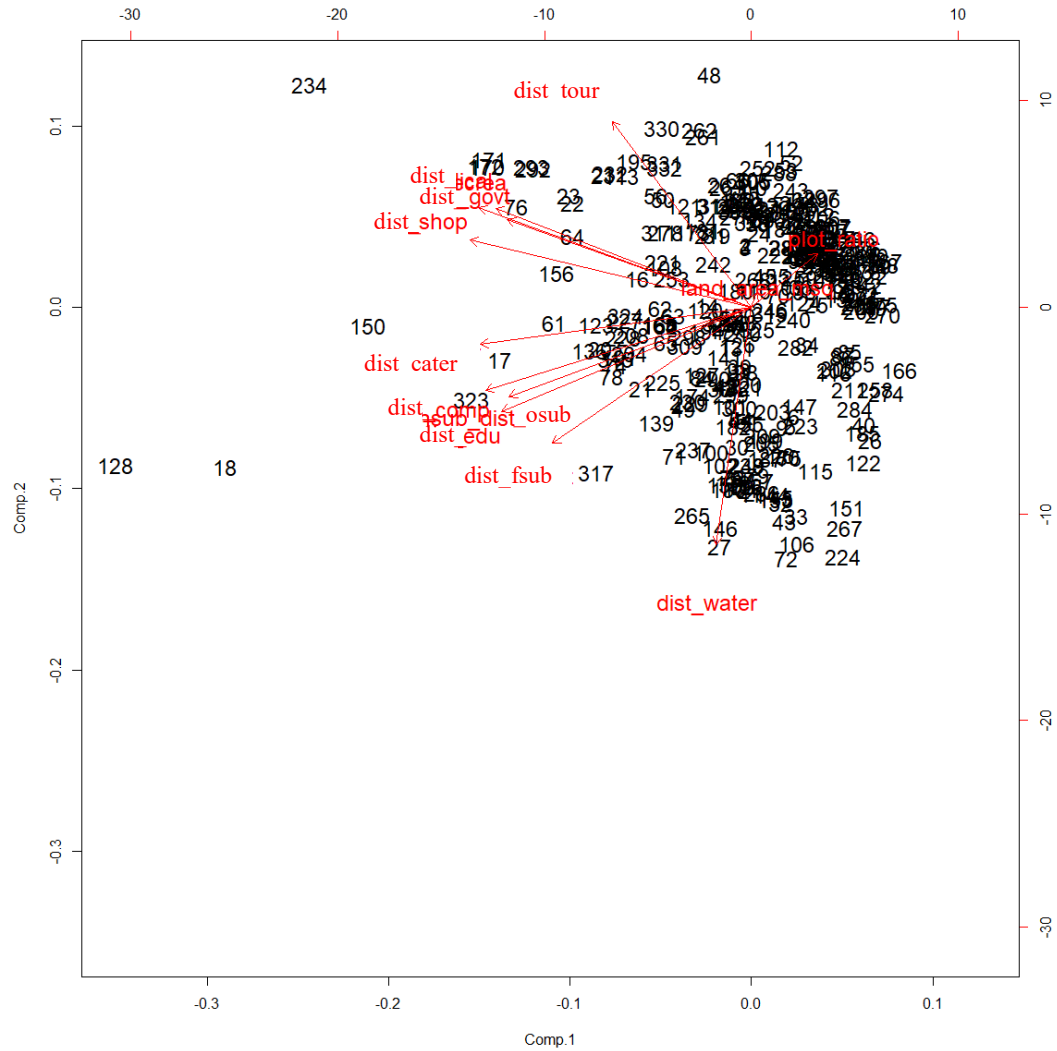


Figure 5.4: Bi-plot of PCA with 13 Variables

Combing the PCA loading factors table and the two-dimension figure we can identify the factors which have less collinearity with each other. At the same time, the predictors selection process should also combine with theories and previous empirical studies about what factors should create impact on the land price.

As discussed in the literature review, the influence of education resource, medical resource, subway system in operation and waterbodies have been detected and explained

across different areas. Further away from these resources, the land price is expected to decrease due to the increasing transportation cost.

For land attributes factors such as land area and plot ratio, although there is no theory of the underlying impact mechanism, empirical works have discovered their influence on land price varying from city to city. In addition, these two factors also have less collinearity with other factors. Therefore, the two land attributes will be kept as independent variables and their effect will be detected and analyzed.

For the locational factor of company, the hypothesis is that large agglomeration of workers created by company clusters will lead to great housing demand, making the real estate development on surrounding land highly profitable. Developers thus will bid higher for the land, resulting in higher transaction price. The relationship between job centers and land price hasn't been studied before. The reason is that although the assumption is well recognized, it is difficult to identify the location of job centers. It is very difficult to acquire location data of employees from government websites. If we adopt the same measurement of distance to individual industrial zone or company with large employment size, we are likely to omit clusters of small companies and enterprises in several concentrated office buildings. And single companies are less stable across time, they often move to other cities or have branches in different areas which will make the assessment of the influence less reliable. With POI data to capture the clusters in the city, the effect of job centers now can be evaluated with data-driven methods. The interaction between job and housing can be also detected and reflected in residential land price.

Many researches have focused on the strong influence of government on economic activities in China. For development of new areas, one of the mostly successfully adopted

strategies is the relocation of bureaus and public services along with new urban construction. The relocation of administration services and resources demonstrates the determination of local government to develop the region. With strong hand controlling policies and resources, the local government will always realize their determinations and bring prosperity to the newly developed areas, manifested in economic growth and pollution increase. In other words, the more administrative facilities clustered in one particular area, the more promising of the area to thrive. For property developers, this regional development pattern indicates proximity to government centers will guarantee decent return of land development and real estate projects due to a predictable population growth in the near future. They are willing to bid higher for land closer to administration centers. To test whether this is the case for Hangzhou city, the variable of distance to government center will be kept.

Similar to subway system under operation, future subway stations also have profound impact on land price. Researches discussed in the literature review also mentioned the expectation of land value premium and more profitable housing development if the land parcel is close to a future subway station. Considering the average real estate development procedure and the construction process of subway, this study set up 3 years as the time length for estimation of anticipation effect of future subway station to land price.

Theoretically speaking, the tourism centers won't create much influence on residential land price, however, in the dataset of POI, tourism category mainly contains natural attractions, parks and other open space resources in the city. This type of environmental attraction will have a certain impact on residential land price. Households are more willing to choose to live in the place where they can enjoy beautiful natural landscape and rich cultural amenities. Developers will then bid higher for the land closer to these centers to meet the

consumers' preferences. The distance to tourism centers is therefore selected as one of the independent variables for the modeling process in the following chapters.

For shopping, catering and entertainment & fitness centers, there is no theory or empirical evidence showing that they have effect on land transaction price. Besides, the VIF calculation combined with bi-plot analysis demonstrated that these factors have strong collinearity issue. Therefore, these three variables are excluded from the following modeling process.

In conclusion, after the comprehensive selection process, three variables are excluded: distance to shopping centers, distance to catering centers and distance to entertainment and fitness. Remaining variables are examined with VIF method again to check the collinearity and the result is listed in Table 5.6:

Table 5.6: VIF Results after Variable Selection

Dependent Variables	VIF
land_area	1.04
plot_ratio	1.13
dist_edu	4.58
dist_cop	6.75
dist_gov	4.48
dist_med	6.49
dist_tour	1.70
dist_water	1.85
dist_osub	5.24
dist_fsub	2.17

The VIF of all the remaining explanatory variables are below 8, which means they are suitable for regression analysis. To further check the correlations among variables, we can conduct PCA again with these eight selected variables. The result is presented in Table 5.7.

Table 5.7: Loadings of Selected Variables in PCA Recalculation

	PC.1	PC.2	PC.3	PC.4	PC.5	PC.6	PC.7	PC.8	PC.9	PC.10
land_area			-0.799	0.558	-0.209					
plot_ratio	0.11	0.165	0.546	0.78		0.207				
dist_edu	-0.409	-0.174		0.154	0.187	-0.125		-0.67	-0.52	
dist_cop	-0.432	-0.112			-0.159	-0.104	-0.48		0.496	-0.531
dist_gov	-0.373	0.285				0.594	0.303	0.2	-0.274	-0.45
dist_med	-0.397	0.324				0.144	0.307	-0.324	0.535	0.476
dist_tour	-0.19	0.55			0.536	-0.455	-0.219	0.318		
dist_water	-0.101	-0.599	-0.126	0.104	0.671	0.227		0.208	0.219	
dist_osub	-0.411	-0.128			-0.32	0.198	-0.403	0.416	-0.242	0.52
dist_fsub	-0.348	-0.253	0.142	0.187	-0.218	-0.509	0.599	0.293		

It is clear that for the variables land_area, plot_ratio, dist_tour and dist_water, they have little correlation with each other. For the rest of the locational factors, although we identified that there are still slight correlations among the factors, it is accepted to keep the selected variables considering the analysis purpose to distinguish differences among factors. More specifically, for the two factors related the city's metro system, in order to look into the differences between current status and anticipation, opened subway stations and future subway stations were separated according to the schedules compared to the land transaction dates. Due to metro planning and system design, the collinearity issue might bear when constructing the two factors. To further check whether the two factors can be separated for the analysis, the scatterplot is provided in the Figure 5.5. From the figure we can see that the two factors didn't show strong correlation. We can conclude that for this study the different influences of the distance to opened metro stations and the distance to future metro stations can be analyzed and incorporated with other impact factors.

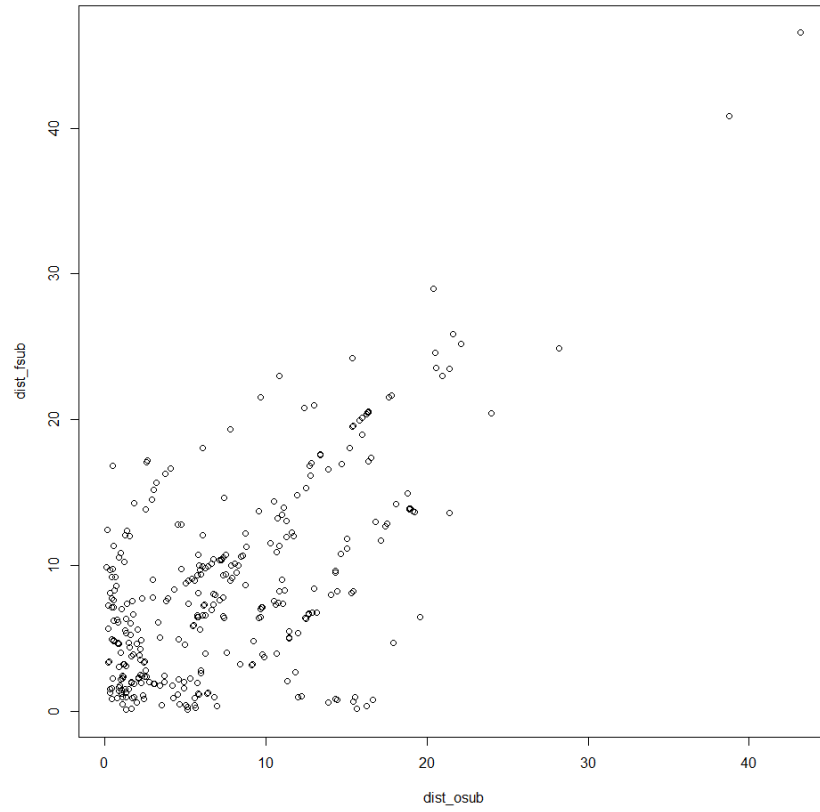


Figure 5.5: Scatterplot of Variables `dist_osub` and `dist_fsub`

CHAPTER 6

OLS ANALYSIS

To test the relationship between the dependent variable and the independent variables, first I used ordinary least square analysis. It is commonly used for parameter estimation with minimization of the sum of the squares of the differences between the observed dependent variable in the given dataset and the prediction by the fitting linear function of independent variables.

The formula of OLS model can be expressed as:

$$\begin{aligned} y_{land\ price} = & \alpha + \beta_1 x_{land_area} + \beta_2 x_{plot_ratio} + \beta_3 x_{dist_edu} + \beta_4 x_{dist_cop} + \beta_5 x_{dist_gov} \\ & + \beta_6 x_{dist_med} + \beta_7 x_{dist_tour} + \beta_8 x_{dist_water} + \beta_9 x_{dist_osub} + \beta_{10} x_{dist_fsub} \\ & + \varepsilon \end{aligned}$$

And the result of coefficients and significance tests are provided in Table 6.1:

Table 6.1: The Result of Coefficients of OLS Model

Impact Factors	OLS Result
land_area	-2.27e-06*** (8.61e-07)
plot_ratio	-0.160*** (0.0447)
dist_edu	0.00194 (0.0132)
dist_cop	-0.0567*** (0.0136)
dist_gov	0.0315 (0.0193)
dist_med	-0.0161 (0.0215)
dist_tour	-0.0359*** (0.0127)
dist_water	-0.0737*** (0.0160)
dist_osub	-0.00280 (0.0106)
dist_fsub	-0.0402*** (0.00659)
constant(α)	10.67*** (0.170)
Observations	332
R-squared	0.569

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

According to the OLS result, distance to the nearest education centers, distance to the nearest government centers, distance to the nearest medical centers and distance to the nearest opened subway stations all have large P-values, implying they have no significant

influence on the land transaction price. Furthermore, the coefficients of distance to the nearest education centers and distance to the nearest government centers are positive, indicating land price will go up as the distance to these centers increase. The results seem to be contradictory to the previous theories and observations. To further verify the significance of factors and provide valid explanation for the relationship between land price and impact factors, GWR is followed with global modeling to explore the possibility of spatial variation.

CHAPTER 7

SPATIAL VARIATION ANALYSIS

7.1 GWR performance

Following with the methodology of GWR discussed in Chapter 4, to set up a GWR model, we need to select an optimal bandwidth for the weighting matrix. Importing the dataset into R programming studio with *sp* and *spgwr* packages¹⁸, we can identify that the best bandwidth for this research is 7.061 kilometers. In other words, the parameter h in the Gaussian kernel function discussed in Chapter 4.4 is determined with this number and the weights of neighboring data points for a local regression are computed relative to this bandwidth and the distance to that local point. For example, a data point at the distance of 7.061 kilometers will be assigned with the weight of $e^{-\frac{1}{2}} = 0.6065$.

The GWR modeling is conducted based on the bandwidth with corresponding kernel functions. The result of local R^2 demonstrated that compared with the global model whose R^2 is 0.57, the R^2 values of the GWR analysis are much higher than that of OLS analysis (see Figure 7.1). This means that the global OLS model can only account for 57 percent of the variance of the dependent variable, while the GWR method represents a significant improvement. In general, A higher R^2 value means that the explanatory variables can explain more variance in land transaction price. It proves that the GWR method produced a more reliable local fitting to the spatial distribution of land price and suggests that the related coefficients are better not assumed to be global (Oliveira, et al., 2014).

¹⁸Information of packages are provided in: <https://www.rdocumentation.org/packages/sp/versions/1.3-1> and <https://www.rdocumentation.org/packages/spgwr/versions/1.3-1>

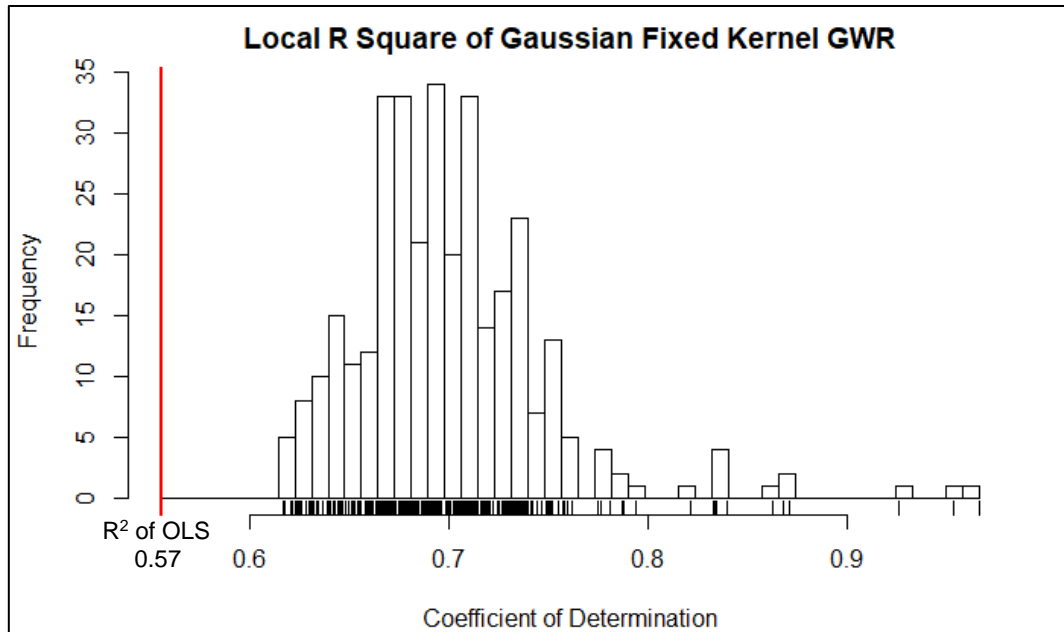
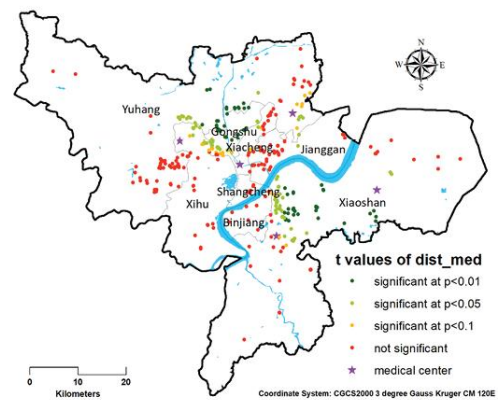
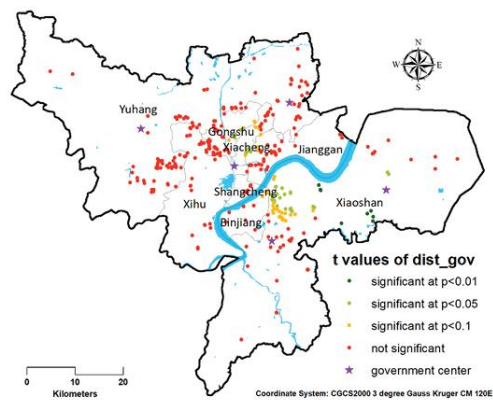
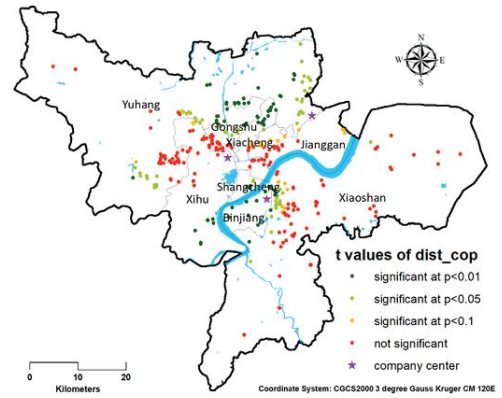
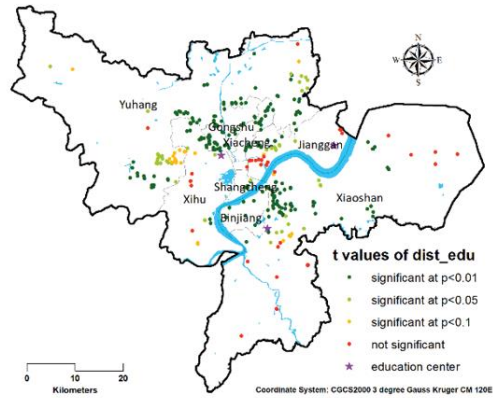
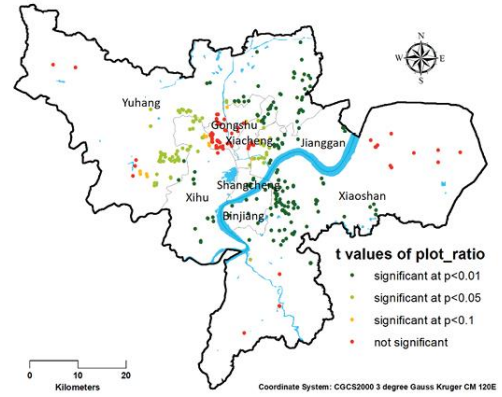
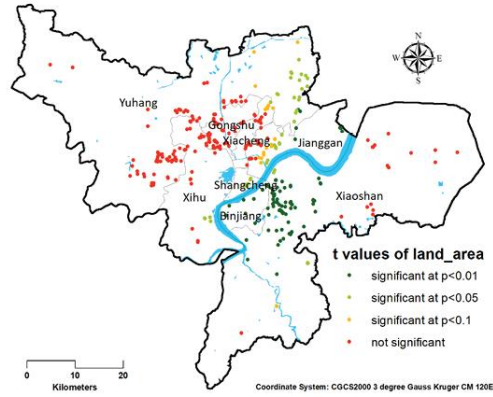


Figure 7.1: Distribution of Local R^2 of GWR Model

7.2 Significance test

The comparison of R^2 of GWR and global model has demonstrated that the GWR model performed much better in explaining the variance over the space. The next step of the GWR modeling is to explore the spatial heterogeneity over the space for each independent variable and further provide explanation for the spatial non-stationary. In addition, for each impact factor, a significance test is conducted by two-tail t-statistics analysis. It is obtained by dividing each local estimate of the regression coefficient by its corresponding local standard error. The result then is transferred into P-value with the degree of freedom of the model for more intuitive interpretation and visualization. The maps of significant test of variables are displayed in Figure 7.2. For the selected ten variables, the distance to the government center are not significant across most of the data points, indicating that agglomerations of bureaus are not influencing the land transaction price. The hypothesis that developers will seek to bid

higher for land closer to this type of clusters for promising return is not supported by the result generated from GWR. The diagnostic shows that for Hangzhou city, during year 2012 -2016, the influence of government on residential land transaction price is not reflected in the distance measurement and proximity analysis. The distance to government clusters is not a suitable predictor for land transaction price change.



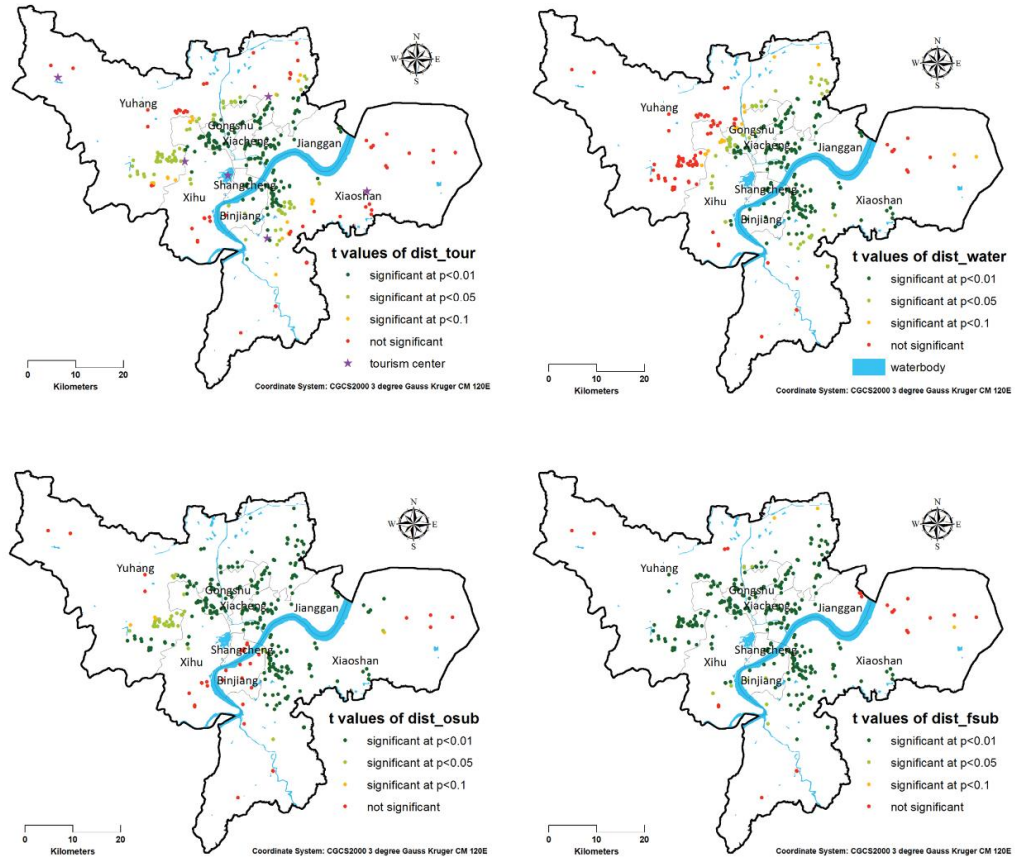


Figure 7.2: Maps of Significant Test for 10 Variables

7.3 Interpretation

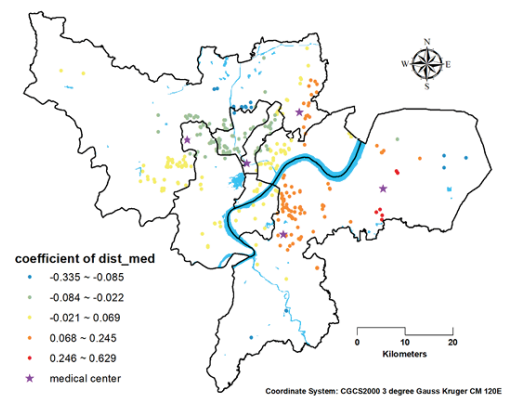
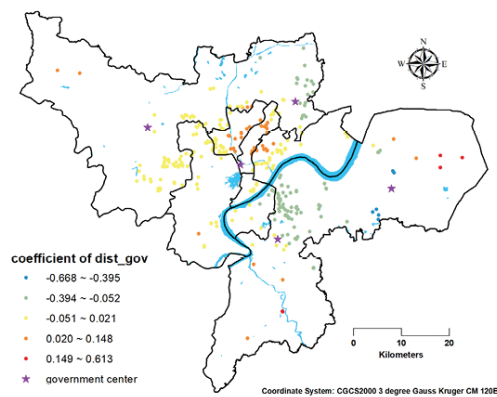
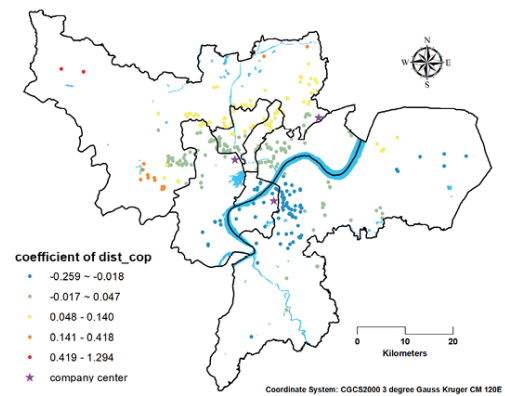
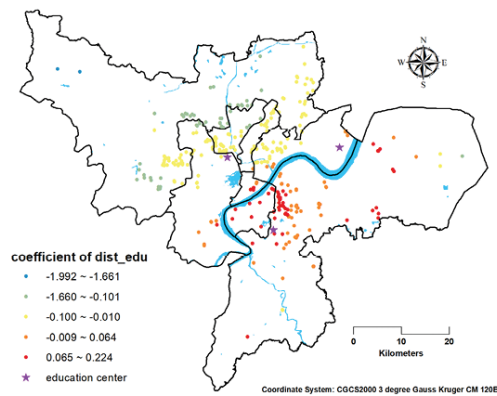
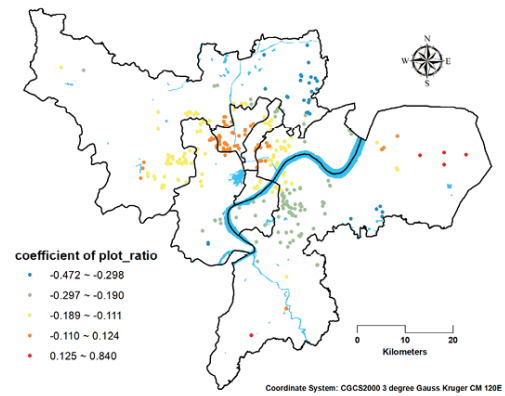
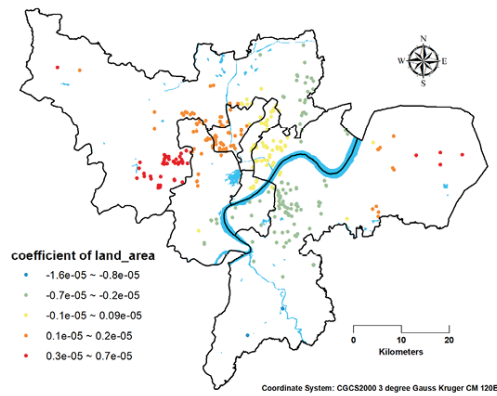
The spatial distribution of coefficients of 10 variables are demonstrated in Figure 7.3. It is clear that the coefficients of all independent variables vary over the city. The spatial heterogeneity is captured by GWR model.

Although the factor of distance to government center is not significant based on the diagnostic in the previous section, the coefficient map is still presented to further prove the unavailability of the predictor: for land parcels located in less developed areas further away from the urban centroid, developers are more likely to offer higher price for land close to

administrative centers for safe return. However, the spatial variation pattern of coefficient didn't provide evidence that closer to the administrative centers, the land price will increase sharper compared to the more developed areas.

Spatial distribution of coefficients of the factors are different, indicating their influence on land price are transmitted through a different mechanism. Many studies have showed that the land appreciation is attributable to the social welfare system such health care services and education. For education centers and medical centers in Hangzhou city, we can identify that the coefficients decrease from southeast to northwest. Further away from the urban core area, the education and medical factors become more influential. This pattern is consistent with findings in previous studies (Colwell & Munneke, 2009). The positive signs of coefficients in the southern part of the city indicate that other factors might be more influential than the two factors. It is possible that the benefit of social welfare system is overwhelmed by attraction of river landscape or convenient commuting hub so that the two social factors don't obtain significant premium for land price in this particular region.

Different from the spatial patterns of the factors of distance to education and medical centers, other impact factors in Hangzhou demonstrate local specific influence patterns different from those of other cities. The discussion will be conducted in the following sections and the influences of factors will be analyzed into groups based on resemblance of mechanism and distribution pattern.



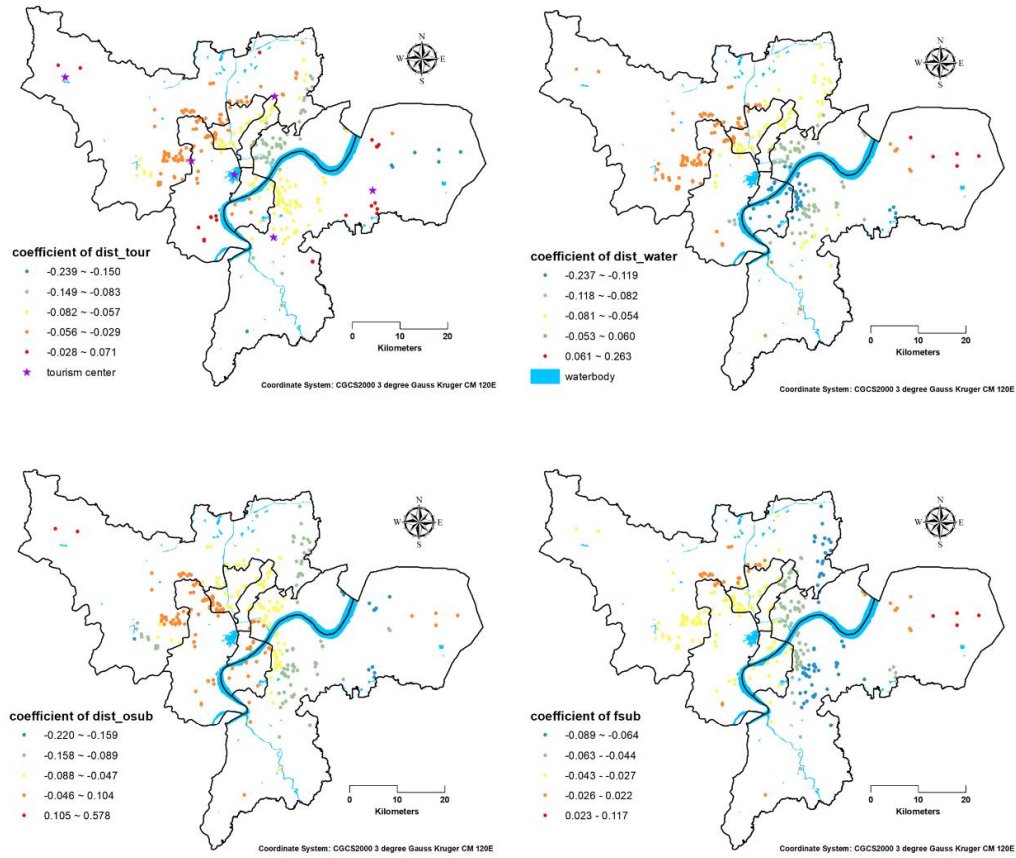


Figure 7.3: Spatial Distribution of Coefficients of Impact Factors

7.3.1 Land attributes: land area and plot ratio

The two major attributes of the land parcel displayed different spatial variation patterns of their influence on the price. For land area factor (see Figure 7.4), its impact on land price is not significant in the north and west of the city, where the coefficients are all positive. For northeastern and southern parts of the city, the coefficients are mostly negative and decreasing in the absolute value from south to north. The positive coefficients in north of Hangzhou (mostly in Gongshu district) were also detected by previous study (Zhang, 2011). The distribution implies that this factor can create stronger impact on land price in the southern

part of the city. When the total area of the land piece increases, the unit bid price will drop more if it is located in the south. Given the fact that these regions are newly developed areas with low population concentration, developers buying larger size of land parcels in the regions may face riskier return to conduct housing projects with inadequate housing demand. It could drive down the demand for land in the regions thus result in lower unit price.

On the other hand, the impact of plot ratio (also known as floor area ratio) presents a concentric spatial pattern which the values of coefficients decrease from the centroid of the city to surrounding areas (Figure 7.5), although some points in the core area show insignificance of the impact of plot ratio to land price. This finding aligns with the previous discussions about floor area ratio influence over the city (Luo, 2007; Zhang, 2011). The comparison with two previous studies demonstrates that the different influence patterns in center and periphery of the city keep consistent over the years. The positive correlation between plot ratio and land price in central city indicates that, in highly developed areas, higher floor area ratio may contribute to higher land premium. The reason can relate to developed infrastructures, denser population and large housing demand. For the less developed areas, developers may face the similar issue with the land area factor: higher floor area ratio may increase the total quantity and the total cost for housing projects built upon the land. In developing area without enough local housing demand, the revenue from real estate development might not be able to cover the cost including land purchase and construction. Therefore, the increase of plot ratio will decrease the unit price more sharply. However, this cost-benefit analysis needs to be further verified by studies on land developers' behaviors. This explanation is to provide one possible framework to identify the driving force behind the spatial heterogeneity.

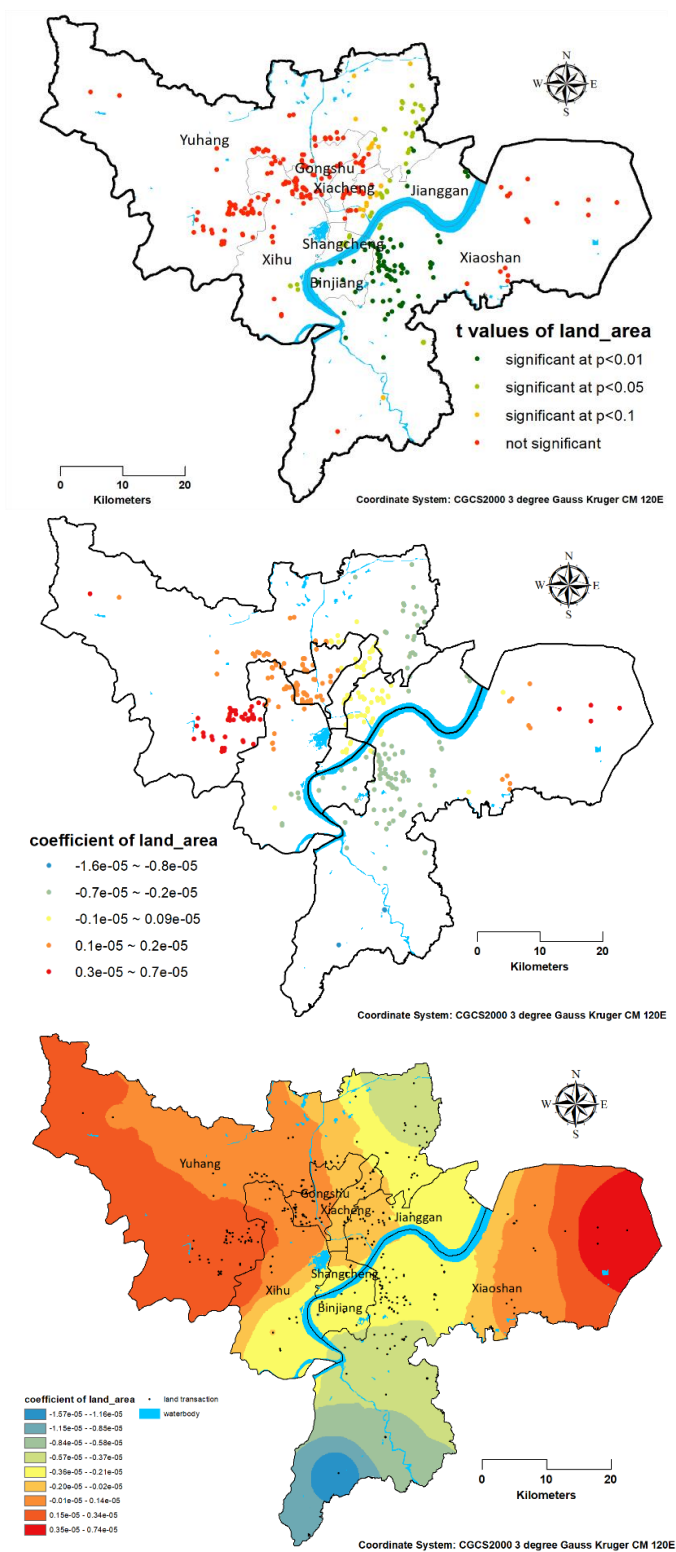


Figure 7.4: GWR Results of the Variable land_area

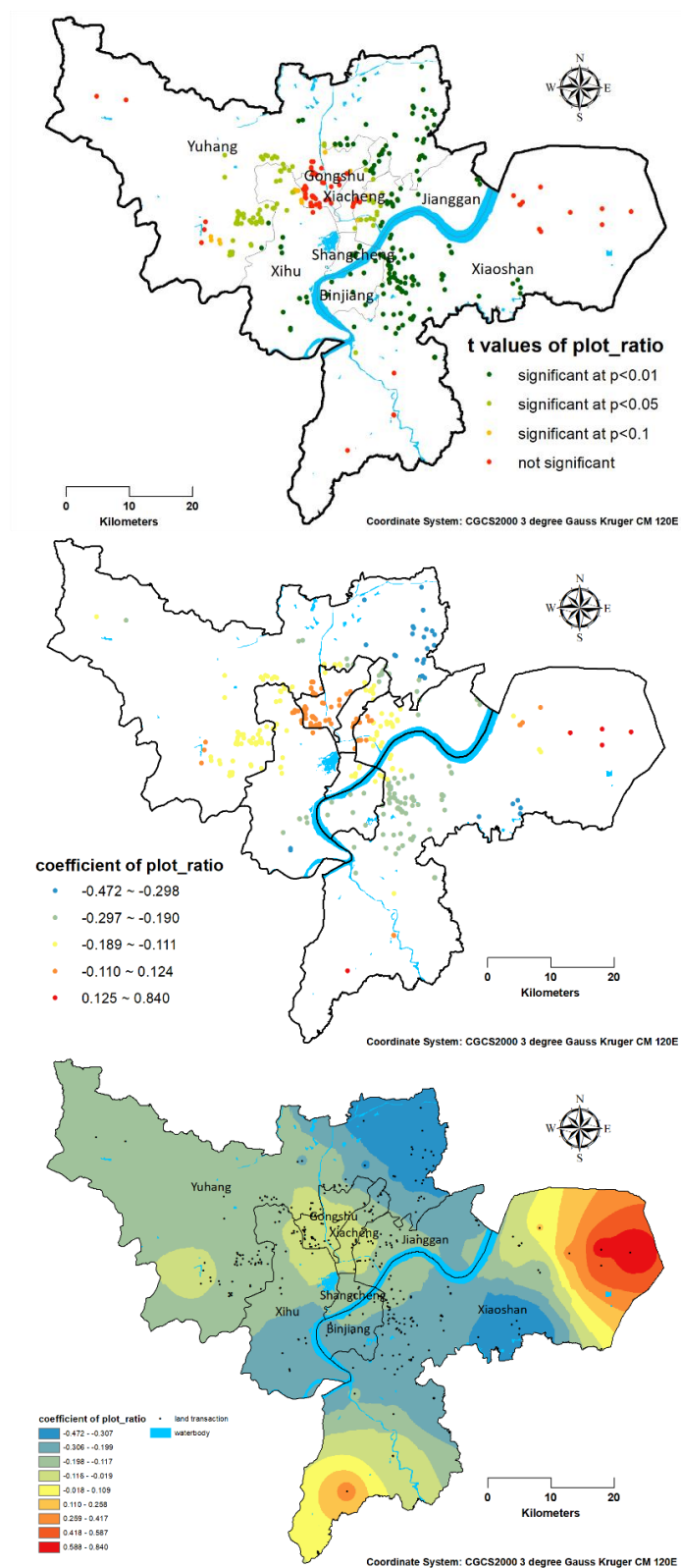


Figure 7.5: GWR Results of the Variable plot_ratio

7.3.2 Pleasant environment: distance to nearest waterbody and distance to nearest tourism center

The definition and specification of variables have been discussed in Chapter 5. These two indicators mainly reflect the attraction of comforting environment and beautiful landscape. Early studies have indicated that negative correlation may exist between land price and lake landscape (Rodriguez & Sirmans, 1994; Bond, Seiler, & Seiler, 2002). Except for very few points where the estimations of coefficients are slightly positive, most of the local estimations demonstrated a negative influence relationship between land price and distance to tourism center and waterbodies. The explanation is that being further away from the natural and cultural resources will lead to the decrease in land price. The spatial distribution of two coefficients shared some level of similarity: the decline of the magnitude of the coefficient of distance to waterbody factor is axisymmetric to the largest waterbody in the city - Qiantang River which divides the city into two parts (Figure 7.6), while the decline of the magnitude of the coefficient of distance to tourism factor is also axisymmetric, along with a southwest-northeast axis (Figure 7.7).

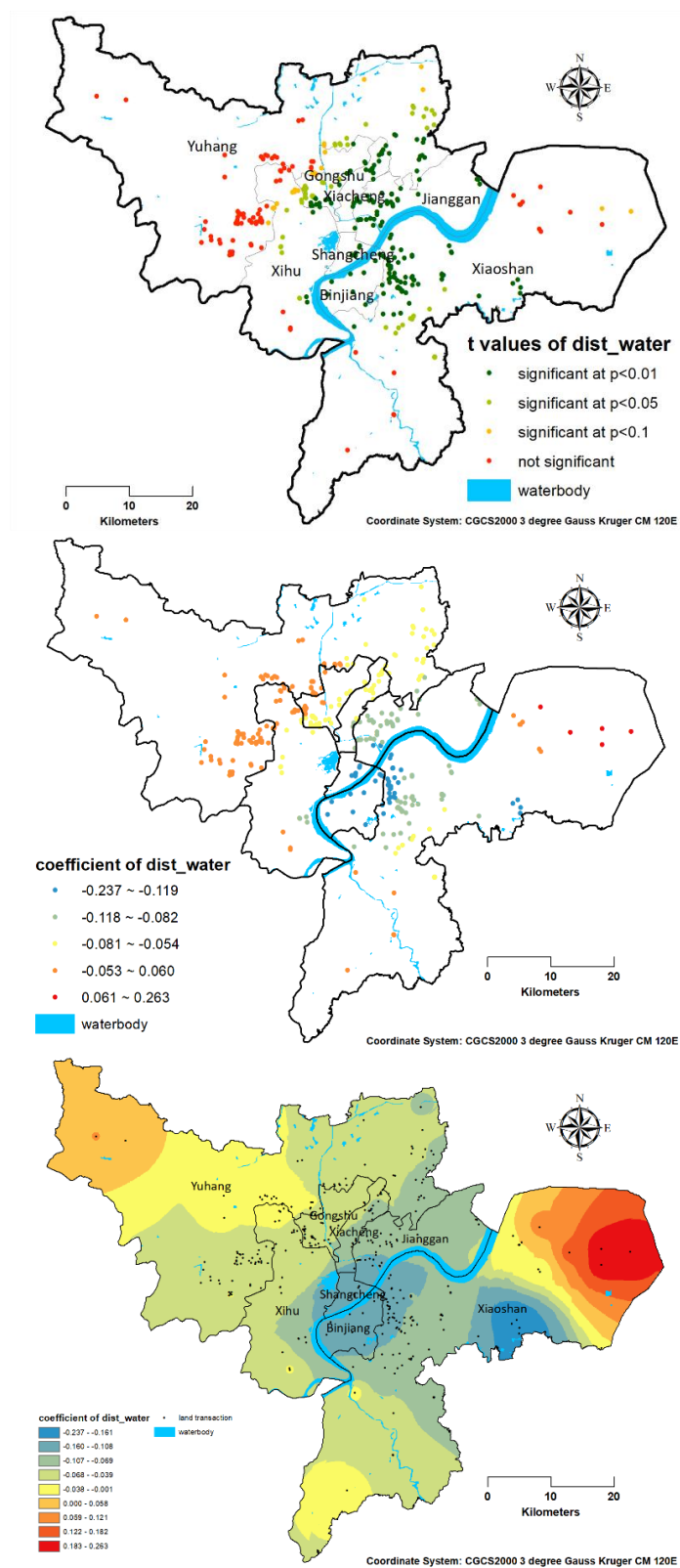


Figure 7.6: GWR Results of the Variable dist_water

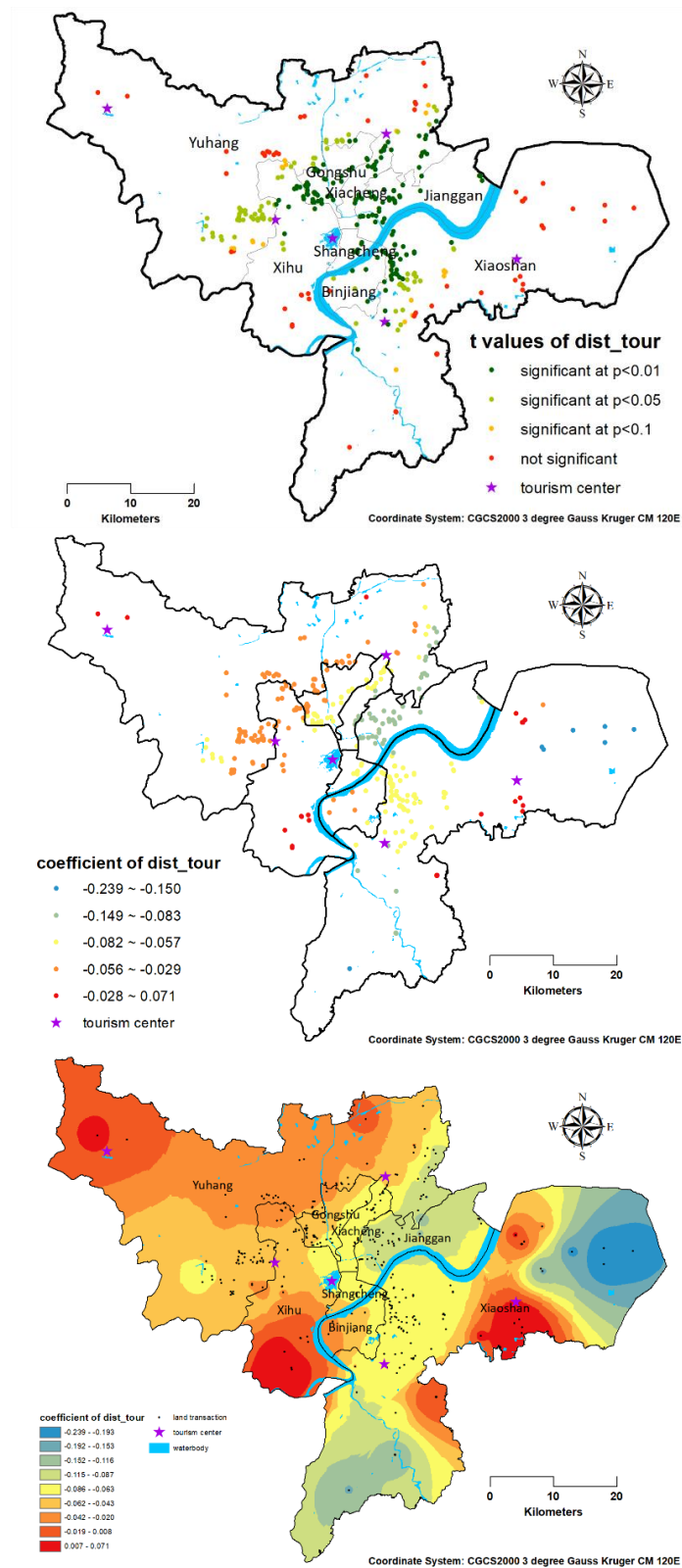


Figure 7.7: GWR Results of the Variable `dist_tour`

The spatial pattern for distance to waterbody should be caused by river landscape with a comfortable and peaceful dwelling environment, where the expansive view of corresponding housing projects is unique and attractive to buyers. The coefficient map of distance to waterbodies also shows that getting closer to the largest river brings greater increase in land bid price. The bidding becomes more competitive at locations very close to the river. In fact, both sides of Qiantang River are famous for their rich neighborhoods. Combining with this fact, it is reasonable to infer from the spatial pattern of the coefficients that developers are bidding higher to get the land parcels nearby in order to meet the demand of wealthier households.

For tourism centers, although the significant test shows that most of land transactions are influenced by the distance to the tourism center, the coefficient map shows that the influence remains mild. It demonstrates that the clusters of natural or cultural tourism resources is correlated to the land price but not creating remarkable impact on the residential land price. Also, there is no clear spatial trend of the coefficient distribution of the factor.

7.3.3 Job-housing balance: distance to the nearest company center

The decay of magnitude of local estimate of coefficients of company centers is starting from the south to the north (Figure 7.8). The spatial pattern indicates that land price is more sensitive to the distance to job centers in the southern part of Hangzhou. For northern part of the city, the influence remains tiny.

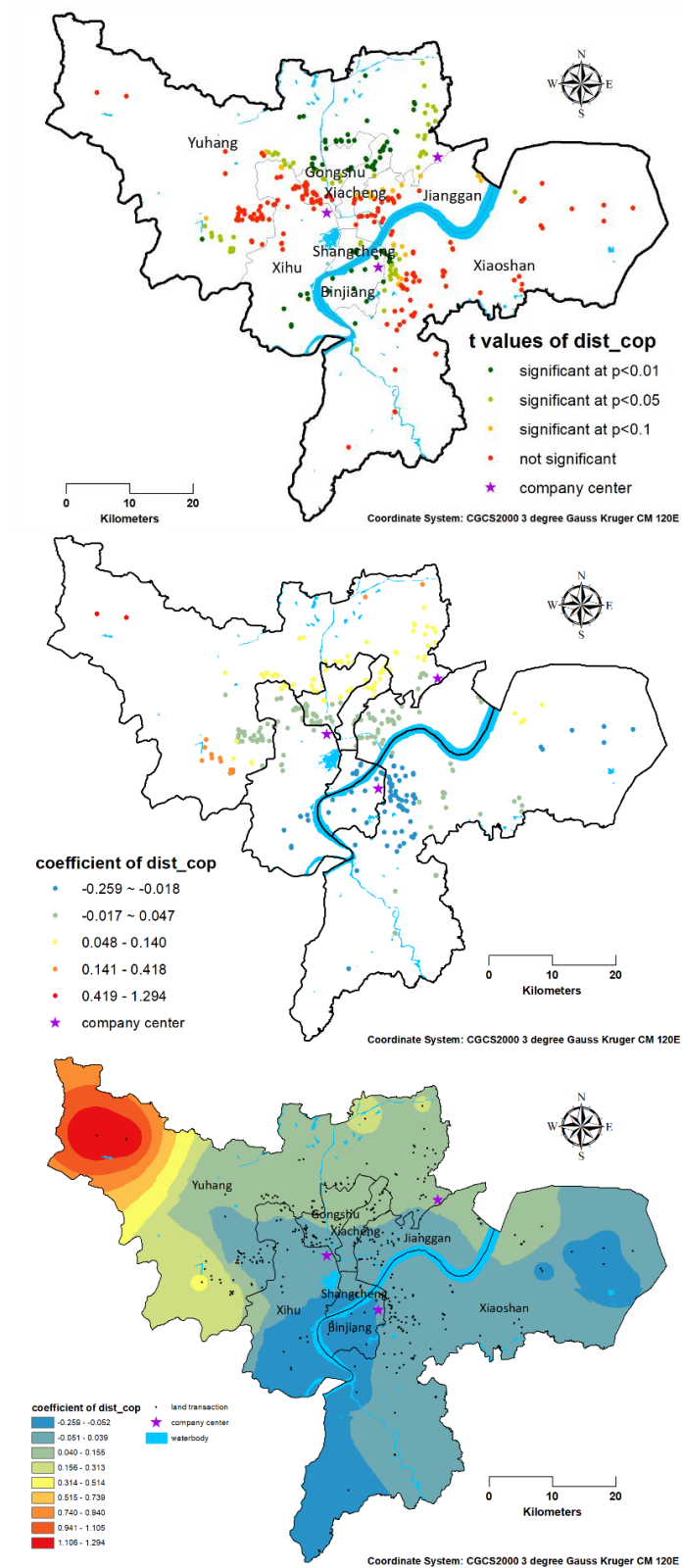


Figure 7.8: GWR Results of the Variable dist_cop

The stronger correlation discovered by GWR in the southern part is noteworthy. The decay pattern is centered around the company center in Binjiang district. Within this area, a small increase in the distance to the company centers will cause a large drop of the land price. The influence becomes weaker further away from this area. The sensitivity for the land price to the distance to the company center can be linked with the special housing demand generated by the industry agglomeration. In this company cluster, many large enterprises set up their business buildings including Alibaba, the biggest e-commerce company in China, research institute center of Huawei and many other IT firms. These business giants create huge number of jobs and rapidly drove up the population in the area. Workers in IT industries receive higher payment and have longer working time. Therefore, they will require closer residence to minimize the commuting cost and balance the job and housing. Housing near the working center will receive higher demand and drive up the price of dwellings. The higher demand for housing close by then could be transferred to a higher demand for the residential land correspondingly, resulting in greater negative impact of distance to company centers to land price. However, the interpretation of the transmission from the housing demand to the land market may need more solid supporting evidences from industry analysis and employee housing survey.

7.3.4 Subway system: stations under operation and stations in the future

The importance of subway system to urban structure has been studied thoroughly by many researches from various aspects. For urban housing market, a subway station nearby will sharply decrease the commuting cost for siting in the location thus allows the household to bid higher for the house (Cardozo, García-Palomares & Gutiérrez, 2012). The same mechanism works for the urban land market, if the land piece is close to a subway station,

households should be willing to pay higher for the dwellings on this land and developers are therefore bidding higher for the land because of the higher revenue they can collect from the buyers.

To analyze the effect of subway system, the comparison between opened subway stations and future stations is conducted. Firstly, the significance test showed that the future stations are more significant across the entire city, almost all the data points are significant at the level of $p < 0.01$. It demonstrated that the correlation of distance to the future subway stations and land transaction price is stronger. The signs of coefficients for both types of variables are mostly negative, which align with the mechanism described above.

The maps in Figure 7.9 and Figure 7.10 clearly showed that the spatial distribution of two types of subway station variables are similar.

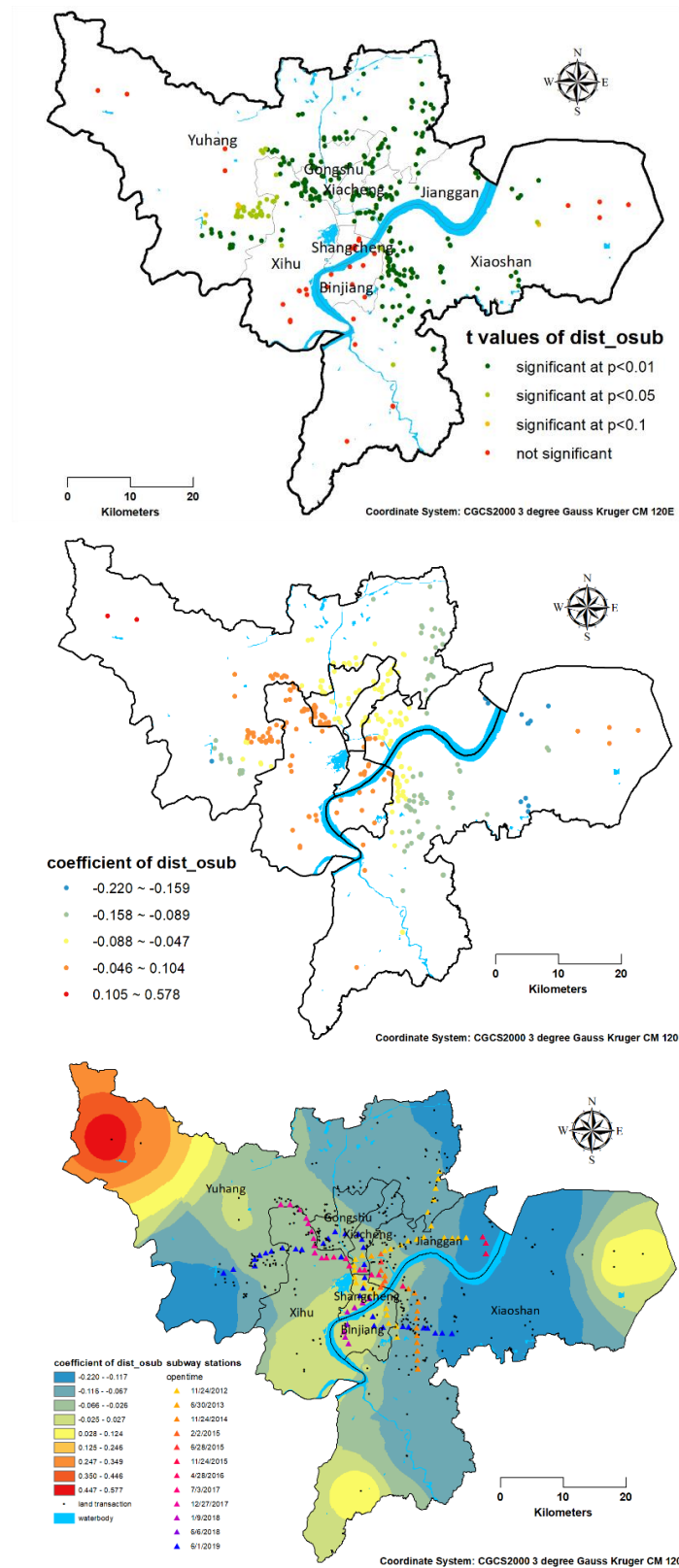


Figure 7.9: GWR Results of the Variable dist_sub

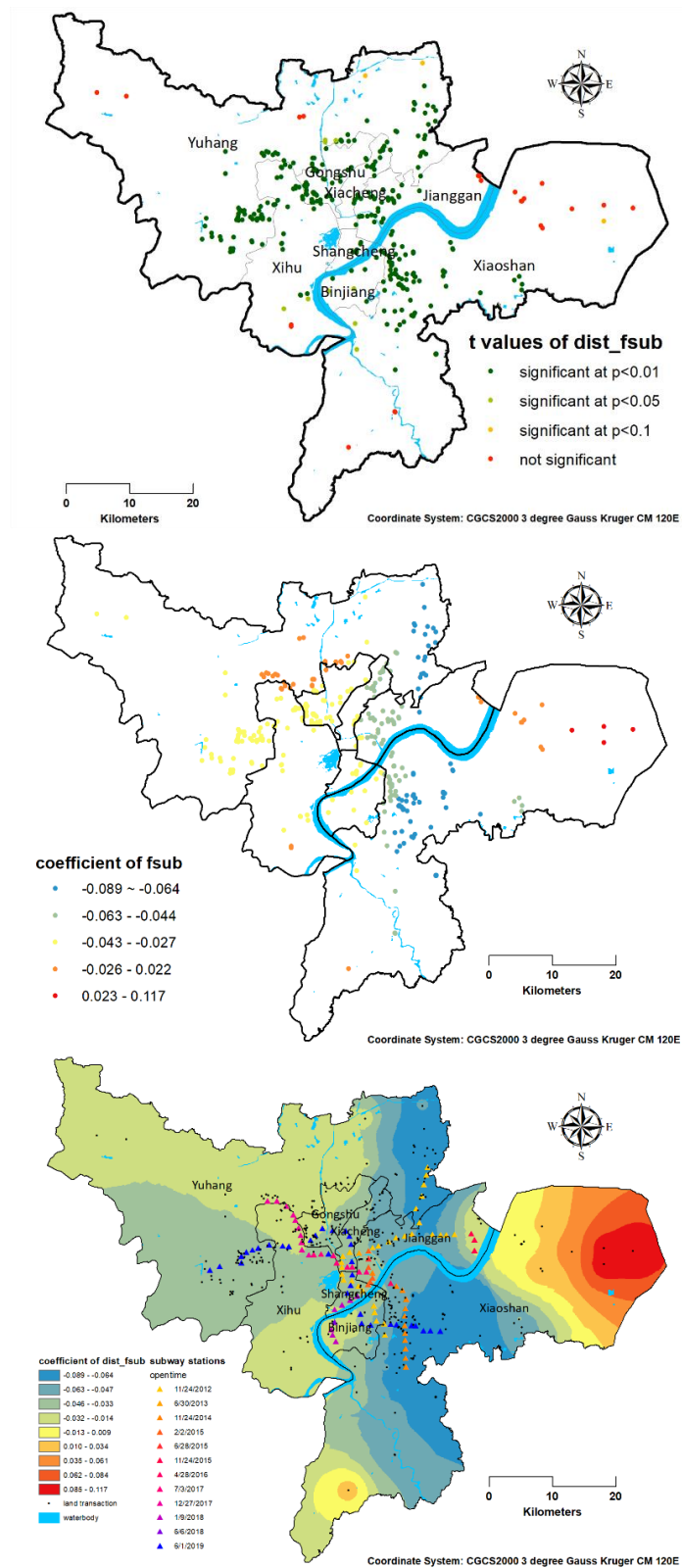


Figure 7.10: GWR Results of the Variable dist_fsub

The pattern of opened subway stations can be described as “terminal effect”: The importance of the opened subway stations to the land price increases as the subway lines extending to periphery area of the city. The magnitude of the coefficients is greater in the terminal part of the subway system. The reason could be that for these undeveloped areas which are located in the periphery of the city, there is no close public transportation substitution for subway. For residents who choose to live in the region, they have to rely on the subway to connect themselves with workplaces and urban facilities located in the core area of the city. Under this scenario, land parcels which are closer to the subway station can attract more consumers to buy the houses built on them. Therefore, the premium potential for the land pieces is higher and developers will bid higher for the land. For the central area of the city, the subway stations don’t create much influence on the land price.

The coefficients distribution of the variable of subway stations opened within 3 years demonstrated similar spatial pattern: Further away from the urban core, the importance of the factor increases. Also, the absolute values of the coefficients increased from the west part of the city to the eastern part. This spatial pattern contradicts to the result of Zhang (2011)’s model. In her analysis, the subway stations which will be opened in the future didn’t demonstrate much significance of influencing the land price, while in this study, most of the data points showed strong significance in terms of distance to future subway stations. Furthermore, the coefficient distribution of this study showed that land parcels closer to the future stations are more sensitive to the distance change, while the previous study found the land transactions further away from stations have greater absolute values of the coefficients.

The similarity of spatial distributions of two types of subway factors is reflected in the overall patterns of coefficients’ change across the city. However, if we look at the magnitude

of the coefficient, we can identify that the coefficients of distance to subway stations which will be opened in the future are smaller than that of the factor of distance to the opened stations. The stronger influence from distance to opened subway stations is not too difficult to understand: Opened subway stations offer actual commuting networks which can be used immediately while future stations offer anticipation of transportation improvement. Opened stations provide a more solid and predictable attraction to housing buyers so the developers will also put more emphasis on the stations under operation rather than on stations under construction or planning.

7.3.5 Summary

Although the spatial distributions of the selected variables are different at certain degree and the underlying mechanisms of interactions with land transaction price are different. The patterns also showed several similarities at the city level.

First, for most impact factors, the data points in the west side of the city didn't show statistical significance, which implied that the bid rent on the land in the west part is not influenced by the factors we selected. One possible explanation is provided for the east-west discrepancy: during 2012 – 2016, the west part of the city was not the target of the government's master plan of urban development. The development strategy of the west side of the city was described as "Quiet and Peaceful"¹⁹ in the planning document in order to protect wetlands and other important natural resources. Therefore, residential land development is limited by the government in these areas. Although there were still land parcels

¹⁹ The spatial development planning can be found in the website of Planning Bureau of Hangzhou City, see: <http://www.hzplanning.gov.cn/Data/HTMLFile/2010-06/423d157e-6fa7-4795-bdd5-d5ecb1148872/fe8f56a4-1db8-4e59-b18b-147cdac3c06a.html?type=%E8%A7%84%E5%88%92%E8%A7%A3%E8%AF%BB%E5%88%97%E8%A1%A8> (in Chinese)

released for developers to bid for residential use, the housing demand was relatively low and the potential for population inflows was also restricted by the policy. Developers bidding for the land in the west were mainly for land stocks. The normalized unit bid price in the area remained relatively low, ranging from 2500 to 4500 RMB per square meter.

Second, the absolute values of coefficients on the north side of the city were smaller than that of the southern part, indicating that the selected impact factors didn't make contribution to the land price change. The differences between coefficients of north and south of the city were caused by the similar reason of the west-east differences: The north part of the city had fewer population to support residential land development and local government didn't put its focus on this region. Developers bid for the land with lower price for stocks for future use. However, as mentioned in the previous section, the educational and medical factors have strong negative influence on the land price in this area. The local history of north Hangzhou city might give helpful information for the reasoning process: the area used to be an industry zone which gathered three major heavy industry factories: the thermal power plant of Hangzhou, Hangzhou oil refinery and Hangzhou steel mill. Residential land released by the government were mostly transferred from industrial land which were in bad conditions. The valuation for the social welfare factors could be viewed as a compensation mechanism for environmental disadvantages. The coefficients of the two impact factors reflected the specific needs in the location which were different from other parts of the city. Therefore, policy making might need to consider social welfare improvement while conducting environmental remediation in the redevelopment process.

In summary, the spatial patterns of the coefficients of different impact factors showed that in different areas of the city, the residential land transaction price is influenced by

different combinations of factors with various driven mechanisms. They also demonstrated that different parts of the city followed their own development paths and developers also responded to the local demands differently. In general, land market in newly developed and less developed areas is more sensitive to the selected impact factors. Admittedly, the GWR model also tells us that this rule doesn't hold for some regions in the city with special local contexts. The variations of significance tests and coefficients proved spatial heterogeneity of impact factors for land price in Hangzhou, and land use policy and urban planning strategy should first try to address these differences and help the city obtain a more integrated development and sustainable growth.

CHAPTER 8

DISCUSSION

8.1 Conclusions

This study analyzed the residential land transaction price and related impact factors in Hangzhou during 2012 – 2016. The spatial heterogeneity identified by the GWR model presented a more detailed relationship between impact factors and land price over the entire city. Instead of applying individual points for locational factors analysis, this research argued that it is the center of points which represents the most attractive location of the corresponding factor that creates influence on the land transaction price. With various types of POI data released from online map service providers, the centers of different urban activities now can be identified with data driven methods. The clusters of urban facilities and services can also help us understand the basic urban structure. K-means clustering and the elbow method were adopted to derive the centers over the space. In addition to the traditional locational factors such as schools, hospitals and waterbodies, the results also characterized the city's job centers which are critical for housing and land markets using the POI data of office buildings and firms. The clusters of education resources identified the university towns in the east side of the city successfully, and the biggest IT industrial agglomeration located in Binjiang district was also captured by the K-means clustering. The result of urban center identification framed the basic urban structure of Hangzhou city during 2012 – 2016 and set a solid foundation for distance calculation and analysis with land parcels.

The variables selection process was conducted by VIF and PCA as well as literature review to derive the independent predictors which are supported by theory and have minimized collinearity. The GWR model built on the selective variables identified the spatial

non-stationarity successfully and displayed better fitting performance compared to the classical OLS model. Spatial distributions of significant tests and coefficients of independent variables revealed local patterns of the influences of particular impact factors. Two subway factors (opened stations and future stations) were proved to have important influence on the land price, indicating developers consider both current benefit of public transportation as well as premium potential in the future. Terminal effect was discovered due to reliance on subway as major commuting mode in the periphery areas of the city. Complementary effect was also analyzed in the north side of the city that proximity to social welfare system (education and medical services) was valued more in order to compensate for the natural disadvantage. Besides regional differences of influence mechanism, at the city level, land transaction price was more sensitive to the selected impact factors in less developed areas than in highly developed areas. Compared to previous studies mainly focusing on the northern part of the city (Luo, 2007; Zhang, 2011), this study identified the spatial heterogeneity in the south of the city for the first time. The model also discovered that southern part of the city is more sensitive to impact factors. Distances to job centers are important to the land transaction price in this region due to the large industry agglomeration and its employee's inflow. With GWR modeling, we are able to disentangle the complexity of influence mechanism of land transaction price over the space and policies can be adapted to local conditions accordingly. For example, local government may need to focus on providing environment improvement on northern areas of Hangzhou, while in the south, more housing projects should be targeted at meeting the demand of labor force in IT industry and floor area ratio of residential land near the company cluster thus can be set up higher to increase the supply for housing at desirable locations.

With technology improvement and institution reform, more data now can be collected and released to be able to be applied with classic theories and achieve new discoveries and breakthroughs. In this study, POI data helped us identify urban centers and corresponding optimal locations of bid rent theory with less subjective judgement. This sort of user generated content (UGC) data will provide richer information in the future and create more raw materials for social science studies. Machine learning has also gained its popularity and been widely applied to cast new light on urban studies (Anderson, 2009; Xu et al., 2011; Gil et al., 2012).

8.2 Limitations

In this section, several limitations of this research are stated and suggestions for improvement are followed accordingly: (1) K-means clustering algorithm employed in this study gives the same weight to all points of interest in their corresponding classifications. With more specified categories and detailed information of POI data, the method can be adjusted with weights of importance of POI combining with hedonic model. (2) The magnitude of coefficients of distances to different types of urban centers is smaller than the result of studies using individual points to calculate the distance influence. The locations of major urban centers only had mild influence on land price. This implies that, other than global centers of the entire city, the local effect also contributes to land transaction differences. Further study can be conducted to incorporate critical local factors with urban centers in order to provide more accurate interpretation of influence mechanism. To better understand locational factors and their impact on land price, more machine learning methods can be applied with diverse data resources. For example, hierarchical clustering combined with POI data can help us understand different levels of urban centers. We can try to find out which level of clustering will create most significant influences to land transaction price using yearly

data and further insight into the reasons behind. (3) This study only focuses on the residential land market. Commercial land and industrial land transactions are also very important for the urban land market and further study can discuss the similarities and differences about three types of land and refine the findings of impact factors analysis. (4) In this research, although the GWR model showed that the locations of government clusters didn't create significant impact on the land price, administrative influence cannot be neglected and requires more in-depth investigation. The impact from government may be transmitted into land market in other forms and need to be detected by other quantification methods. (5) One of the important assumptions of this study is that the basic urban structure of Hangzhou didn't change during 2012 – 2016. This is due to data availability and policy changes. The uneven distribution of land transactions over the space and time further adds to the difficulty of developing a more precise model to characterize the spatial non-stationarity.

With data accessibility improvement in the future, the modeling process can be separated year by year. The temporal and spatial differences of land price and its influential factors can be captured simultaneously and provide explanation of interactions between land market and urban growth. While existing data and methods may not be sufficient to obtain a comprehensive analysis of spatial heterogeneity of residential land price. The study has made a progress towards the goal.

REFERENCES

- Allison, P. D. (1999). *Multiple regression: A primer*. Pine Forge Press.
- Alonso, W. (1964). *Location and land use*. Cambridge: Harvard University Press.
- Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41(3), 359-364.
- Bond, M. T., Seiler, V. L., & Seiler, M. J. (2002). Residential real estate prices: A room with a view. *The Journal of Real Estate Research*, 23(1), 129-137.
- Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4), 281-298.
- Butler, R. V. (1982). The specification of hedonic indexes for urban housing. *Land Economics*, 58(1), 96-108.
- Cao, T., Huang, K., Li, J., Dong, P., & Wang, Y. (2013). 基于 GWR 的南京市住宅地价空间分异及演变 [Research on spatial variation and evolution of residential land price in Nanjing based on GWR Model]. *Geographical Research/地理研究*, 32(12), 2324-2333.
- Cardozo, O. D., García-Palomares, J. C., & Gutiérrez, J. (2012). Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, 34, 548-558.
- Chau, K. W., & Chin, T. L. (2003). A critical review of literature on the hedonic price model. *International Journal for Housing Science and Its Applications*, 27(2), 145-165.
- Colwell, P. F., & Munneke, H. J. (2009). Directional land value gradients. *The Journal of Real Estate Finance and Economics*, 39(1), 1-23.
- Cui N., Feng C., & Song Y. (2017). 北京市居住用地出让价格的空间格局及影响因素 [Spatial pattern of residential land parcels and determinants of residential land price in Beijing since 2004]. *Acta Geographica Sinica/地理学报*, 72(6), 1049-1062.
- Dziauddin, M. F. (2019). Estimating land value uplift around light rail transit stations in Greater Kuala Lumpur: An empirical study based on geographically weighted regression (GWR). *Research in Transportation Economics*.
- Fujita, M., & Thisse, J. F. (2013). *Economics of agglomeration: cities, industrial location, and globalization*. Cambridge university press.

- Gao, J., Chen, J., & Su, X. (2014). 2001-2010 年南京市土地出让价格的影响因素 [Influencing factors of land price in Nanjing Proper during 2001-2010]. *Progress in Geography/地理科学进展*, 33(2), 211-221.
- Gil, J., Beirão, J. N., Montenegro, N., & Duarte, J. P. (2012). On the discovery of urban typologies: data mining the many dimensions of urban form. *Urban morphology*, 16(1), 27.
- Glumac, B., Herrera-Gomez, M., & Licheron, J. (2019). A hedonic urban land price index. *Land Use Policy*, 81, 802-812.
- Goffette-Nagot, F., Reginster, I., & Thomas, I. (2011). Spatial analysis of residential land prices in Belgium: accessibility, linguistic border, and environmental amenities. *Regional Studies*, 45(9), 1253-1268.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6), 417.
- Hu, S., Yang, S., Li, W., Zhang, C., & Xu, F. (2016). Spatially non-stationary relationships between urban residential land price and impact factors in Wuhan city, China. *Applied Geography*, 68, 48-56.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.
- Jones, M., & Reed, R. G. (2018). Open space amenities and residential land use: An Australian perspective. *Land use policy*, 75, 1-10.
- Kan, B., Pu, L., Xu, C., Huang, S., Zhu, M., Huang, S., & Xie, Z. (2019). 基于 GWR 模型的南京主城区住宅地价空间异质性驱动因素研究 [Driving Factors on the Spatial Heterogeneity of Residential Land Price in Downtown Nanjing Based on GWR Model.] *Economic Geography/经济地理*, 1-11. Retrieved from <http://kns.cnki.net/kcms/detail/43.1126.K.20181229.1125.014.html>.
- Kanasugi, H., & Ushijima, K. (2018). The impact of a high-speed railway on residential land prices. *Papers in Regional Science*, 97(4), 1305-1335.
- Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6), 441-458.
- Luo, G. (2007). 基于 GWR 模型的城市住宅地价空间结构研究 [Spatial Structure of Urban Housing Land Prices Based on GWR Model] (Doctoral dissertation, Zhejiang University, Hangzhou, China).

- Lv, P., & Zhen, H. (2010). 基于 GWR 模型的北京市住宅用地价格影响因素及其空间规律研究 [Affecting factors research of Beijing residential land price based on GWR model]. *Economic Geography/经济地理*, 3, 472-478.
- Mulley, C., & Tsai, C. H. P. (2016). When and how much does new transport infrastructure add to property values? Evidence from the bus rapid transit system in Sydney, Australia. *Transport Policy*, 51, 15-23.
- Nie C., Wen H., & Fan X. (2010). 城市轨道交通对房地产增值的时空效应 [The spacial and temporal effect on property value increment with the development of urban rapid rail transit: An empirical research]. *Geographical Research/地理研究*, 29(5), 801-810.
- Nilsson, P. (2014). Natural amenities in urban space—A geographically weighted regression approach. *Landscape and Urban Planning*, 121, 45-54.
- Oliveira, S., Pereira, J. M., San-Miguel-Ayanz, J., & Lourenço, L. (2014). Exploring the spatial patterns of fire density in Southern Europe using Geographically Weighted Regression. *Applied Geography*, 51, 143-157.
- Philip, G. M., & Watson, D. F. (1982). A precise method for determining contoured surfaces. *The APPEA Journal*, 22(1), 205-212.
- Puu, T. (2012). *Mathematical location and land use theory: an introduction*. Springer Science & Business Media.
- Qu, S., Hu, S., Yang, S., & Li, Q. (2018). 城市住宅地价影响因素的定量识别与时空异质性——以武汉市为例 [Quantitative evaluation of the impacts of driving factors on urban residential land price and analysis of their spatio-temporal heterogeneity: A case study of Wuhan City]. *Progress In Geography/地理科学进展*, 37(10), 1371- 1380.
- Rodriguez, M., & Sirmans, C. F. (1994). Quantifying the value of a view in single-family housing markets. *Appraisal Journal*, 62, 600-600.
- Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1), 34-55.
- Sui X., Wu W., & Zhou S., et al. (2015). 都市新区住宅地价空间异质性驱动因素研究——基于空间扩展模型和 GWR 模型的对比 [Drive pattern on the spatial heterogeneity of residential land price in urban district: A comparison of spatial expansion method and GWR model]. *Scientia Geographica Sinica/地理科学*, 35(6): 683-689.
- Tamesue, K., & Tsutsumi, M. (2013). Geographically weighted regression approach for origindestination flows. In *VII World Conference of the Spatial Econometrics Association*.

Watson, D. F. (1985). A refinement of inverse distance weighted interpolation. *Geoprocessing*, 2, 315-327.

Xu, K., Kong, C., Li, J., Zhang, L., & Wu, C. (2011). Suitability evaluation of urban construction land based on geo-environmental factors of Hangzhou, China. *Computers & Geosciences*, 37(8), 992-1002.

Yuan, N. J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., & Xiong, H. (2015). Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27(3), 712-725.

Zhang, J. (2011). 基于 GWR 模型的城市住宅地价的空间分异研究-以杭州市为例 [Research on the Spatial Variation of the Urban Housing Land Prices Based on Geographically Weighted Regression Model-A Case Study of Hangzhou] (Master's thesis, Zhejiang University, Hangzhou, China).